Virus discovery using current and novel methods

Barbara Franziska Mühlemann

Jesus College

This dissertation is submitted for the degree of Doctor of Philosophy

University of Cambridge, June 2019

Supervised by Dr. Terry C. Jones and Prof. Derek J. Smith

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface of the relevant chapters and specified in the text. It is not substantially the same as any that I have submitted, or is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the Degree Committee for the Faculty of Biology.

Cambridge, June 2019

Barbara Franziska Mühlemann

SUMMARY

Virus discovery using current and novel methods

Barbara Franziska Mühlemann

Next Generation Sequencing (NGS) technology allows researchers to sequence genetic material from a wide range of sources, including patient and environmental samples, and ancient remains. The recovery of viruses from such datasets can provide insights into the diversity and evolution of both novel and already known viruses. This thesis focuses on two aspects of virus discovery in NGS datasets.

In the first part of this thesis, I present ancient viral sequences from hepatitis B virus, human parvovirus B19, and variola virus. The sequences were recovered from NGS datasets from individuals living in Eurasia between ~150 to ~31,630 years ago, using standard sequence matching tools. The data show the past existence of viruses similar to variants circulating today. The sequences reveal a complexity of virus evolution that is not evident when considering modern sequences alone, including revised substitution rates and most recent common ancestor dates, as well as geographic movement and extinction of strains.

The identification of viral sequences in NGS datasets relies heavily on sequence-based matching of unknown sequences to a database of known sequences. Comparisons are usually done at the nucleotide or amino acid level. However, those methods only work well on sequences closely related to those already present in the database. With the aim of identifying more diverged viral sequences, in the second part of this thesis, I present an algorithm to compare sequences based on predicted structural features, such as secondary structures and conserved amino acids. The algorithm is modelled after the music-matching algorithm 'Shazam'. While initial results of the algorithm are somewhat encouraging, problems remain, in particular with the identification of adequate structural features. Identifying highly diverged viral sequences is thus still a challenging problem, hopefully to be solved in the future.

In memory of
Beatrice Mühlemann-Vollmeier

ACKNOWLEDGEMENTS

I am very grateful to have been supervised by Terry Jones and Derek Smith. Their approach to science, their curiosity, and thoroughness were truly inspiring. I could not have wished for better mentors and to receive better training as a PhD student. I cannot thank Derek enough for giving me the opportunity to work towards a PhD in his group, and for his generosity and support in everything I chose to do. Terry has been a constant source for advice and sounding board for ideas. He is extremely generous with his time and knowledge, and his directness and humour made it a pleasure to work together. Terry taught me how to write code, about the importance of testing, and being strict about code quality. Thank you so much for the advice and pep-talks, lunches, dinners, coffees, and much more!

The work on ancient viruses would not have been possible without the collaboration with Eske Willerslev and his team at the Centre for GeoGenetics in Copenhagen and Cambridge. As a result, we were very fortunate to be trusted to work with truly amazing data. I am hugely grateful to Eske Willerslev, Lasse Vinner, Morten Allentoft, Ashot Margaryan, Peter de Barros Damgaard, Simon Rasmussen, Hugh McColl, Martin Sikora, Maanasa Raghavan, Hannes Schroeder, and Constanza de la Fuente Castro, who have spent countless hours of work, and money to assemble and sequence the datasets analysed in the first part of this thesis. Lasse Vinner also performed the targeted capture work specifically for the papers on ancient hepatitis B virus, human parvovirus B19, and variola virus and designed the capture probes together with Anders Hansen. Furthermore, I am very grateful to Eppie Jones and Turi King, who have provided additional datasets, and to Andrea Manica for putting us in contact.

I am grateful to Christian Drosten and Ron Fouchier, who both have provided valuable comments on various aspects of the work presented in this thesis over the years.

Many thanks to Dieter Glebe, for answering all our hepatitis B virus related questions, his support on the hepatitis B virus paper, and his continued enthusiasm. Geoffrey Smith and Gerd Sutter generously shared their knowledge on variola virus and vaccinia virus gene functions – thank you!

I am grateful to the archaeologists and curators who have provided samples from ancient individuals to members of the Centre for GeoGenetics for sequencing. They are: Kristian Kristiansen, Karl-Göran Sjögren, Helene Wilhelmson, Andrzej Weber, Irina Shevnina, Andrey Logvin, Emma Usmanova, Irina Panyushkina, Bazartseren Boldgiv, Tsevel Bazartseren, Kadicha Tashbaeva, Victor Merz, Nina Lau, Václav

Smrčka, Dmitry Voyakin, Egor Kitov, Andrey Epimakhov, Dalia Pokutta, Magdolna Vicze, Douglas Price, Vyacheslav Moiseyev, Vladimir I. Bazaliiskii, Martyna Molak-Tomsia, Jette Arneborg, Wieslaw Bogdanowicz, Ceri Falys, Mikhail Sablin, Sabine Sten, Niels Lynnerup, Lisa Mariann Strand, Jan Bill, Alexandra Buzhilova, Tamara Pushkina, Valeri Khartanovich, Marie Louise Schjellerup Jørkov, and Palle Østergaard Sørensen. Learning about the details of what is known about the lives of those individuals, in discussions in person and over email, has added another dimension to the study of ancient viruses that I greatly enjoyed.

My advisors, Jim Kaufman and Andrea Manica, have provided comments and suggestions at various stages of this PhD, and I am grateful for their advice and stories from their own careers.

Thanks to my co-workers in Cambridge, Sam Wilks, David Pattinson, David Burke, Judy Fonville, Leah Katzelnick, Ana Mosterin-Höpping, Daniel Fabian, Blake Hauser, Sarah James, Poppy Roth, Pete Thomas-McEwen, and Longzhu Shen for the chats, the shared lunches, dinners, and coffee breaks.

I would like to thank my friends, Ariane Nikpur, Wan Fong Woo, Dana Usher, Talitha Veith, Anisa Lulo, Noémie Schegg, Julia Hättenschwiler, Albert Perez-Riba, Tarjinder Singh, Michelle Rigozzi, Rahel Frei, Viola Frei, and Wei Li. The last four years would have been far less colourful without them!

Finally, a line in a thesis will never do justice to the lifetime of love and support that I received from my parents, Beatrice and Kaspar Mühlemann. From the bottom of my heart, thank you!

PUBLICATIONS

Versions of chapters 2, 3, and 4 have previously been published, or are currently in review, as the following papers (* authors contributed equally):

Barbara Mühlemann*, Terry C. Jones*, Peter de Barros Damgaard*, Morten E. Allentoft*, Irina Shevnina, Andrey Logvin, Emma Usmanova, Irina P. Panyushkina, Bazartseren Boldgiv, Tsevel Bazartseren, Kadicha Tashbaeva, Victor Merz, Nina Lau, Vàclav Smrčka, Dmitry Voyakin, Egor Kitov, Andrey Epimakhov, Dalia Pokutta, Magdolna Vicze, T. Douglas Price, Vyacheslav Moiseyev, Anders J. Hansen, Ludovic Orlando, Simon Rasmussen, Martin Sikora, Lasse Vinner, Albert D. M. E. Osterhaus, Derek J. Smith, Dieter Glebe, Ron A. M. Fouchier, Christian Drosten, Karl-Göran Sjögren, Kristian Kristiansen, Eske Willerslev. *Ancient Hepatitis B viruses from the Bronze Age to the Medieval period.* Nature, 557, 418-423 (2018).

Barbara Mühlemann*, Ashot Margaryan*, Peter de Barros Damgaard*, Morten E. Allentoft*, Lasse Vinner, Anders J. Hansen, Andrzej Weber, Vladimir I. Bazaliiskii, Martyna Molak, Jette Arneborg, Wieslaw Bogdanowicz, Ceri Falys, Mikhail Sablin, Vàclav Smrčka, Sabine Sten, Kadicha Tashbaeva, Niels Lynnerup, Martin Sikora, Derek J. Smith, Ron A. M. Fouchier, Christian Drosten, Karl-Göran Sjögren, Kristian Kristiansen, Eske Willerslev, Terry C. Jones. *Ancient human parvovirus B19 in Eurasia reveals its long-term association with humans.* PNAS, 115(29), 7557-7562, (2018).

Barbara Mühlemann*, Martin Sikora*, Lasse Vinner*, Ashot Margaryan, Helene Wilhelmson, Constanza de la Fuente Castro, Morten E. Allentoft, Peter de Barros Damgaard, Anders Johannes Hansen, Sofie Holtsmark Nielsen, Lisa Mariann Strand, Jan Bill, Alexandra Buzhilova, Tamara Pushkina, Ceri Falys, Valeri Khartanovich, Vyacheslav Moiseyev, Marie Louise Schjellerup Jørkov, Palle Østergaard Sørensen, Hannes Schroeder, Gerd Sutter, Geoffrey L. Smith, Christian Drosten, Ron A. M. Fouchier, Derek J. Smith, Terry C. Jones, Eske Willerslev. *Diverse variola virus lineages circulated in northern Europe during the Viking Age*. In review at Science, February 2019.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Viruses are ubiquitous and diverse. Virus particles have been found in all environments examined so far, outnumber cells by a factor of 10 – 100 in most well-studied habitats [1], and infect all forms of life [2]. The diversity of viruses is illustrated by the differences in the type and size of viral genomes, virion morphology and replication strategies: virus genomes consist of either RNA or DNA, can be single or double-stranded, linear or circular, multi- or monopartite, and positive or negative sense [2]. Viral morphologies are also highly diverse, including icosahedral, spherical, bottle-shaped, rod-shaped, filamentous, pleomorphic, turreted, and bacilliform [2]. Finally, viruses employ at least seven different replication strategies, which provides the basis for the Baltimore classification [2].

Viruses play an important role in ecosystem and evolutionary processes [1,3,4]. Up to 40% of the marine prokaryote community is killed by viruses every day [3]. Organic molecules released from the viral lysis of those organisms in turn stimulate further growth of phyto- and zooplankton [3]. Viruses also influence the evolution of their hosts [1, 4]. Hosts have multiple lines of defence against infections by pathogens (for example the adaptive and innate immune systems, and programmed cell death), and viruses have evolved ways to evade those mechanisms (including inhibitors of apoptosis or signalling pathways leading to the production of interferons, as well as proteins directly inhibiting interferons, cytokines, and chemokines), suggesting that viruses and their hosts may at least in part be driving each others evolution [1, 5–7]. Viruses are an important mechanism for the transfer of genes between organisms, some of which may be beneficial to the recipient [4]. There are a number of examples where a gene that originally stems from a virus is now fulfilling an important role in a cellular organism [4, 8]. Probably the best-known is the case of syncytin, a protein that originated from the envelope protein of the human endogenous retrovirus W, and which is essential for placental morphogenesis in mammals [8].

Viruses are also the cause of great devastation in humans and other organisms. While some viruses usually cause mild or self-limiting disease, such as viral gastroenteritis caused by norovirus or common cold caused by rhinoviruses, others can cause severe

disease or death, for example infections by ebola virus (with an average case fatality rate of 78% [9]), variola virus (where the case fatality rate varied between 1 – 30% [10]), and rabies viruses (which is almost always fatal after the onset of clinical symptoms [11]). In animals, viruses have been the cause of die-offs, such as those caused by highly pathogenic avian influenza virus in chickens and turkey [12, 13], or highly virulent strains of African swine fever virus in wild and domestic pigs, where case fatality rates are almost 100% [14]. Using a dataset of 335 disease emergence events between 1940 and 2004[1], Jones *et al.*, (2008) estimated that a quarter of all emerging infectious diseases in humans are caused by viruses [15]. Disease emergence is correlated with human population density, suggesting that it is driven by anthropogenic and demographic changes such as human population growth, deforestation, increased mobility, and urbanisation [15]. These drivers are expected to remain constant or increase in the future, possibly resulting in increased disease emergence [16–19].

In the past, the discovery of novel viruses relied on techniques such as filtration, cultivation, and electron microscopy [20]. In particular, cultivation has been the gold standard for virus discovery [21]. However, many viruses cannot be grown in culture [22], which limits the use of tissue culture for virus discovery. The development of Next Generation Sequencing (NGS) technology has made the sequencing of genetic material fast and cheap. Since 2008, when NGS technology first came in use, the cost to sequence one megabase of DNA has decreased almost 10^5-fold [23]. Among other applications, NGS technology allows the sequencing and subsequent analysis of all genetic material present in a sample, a so-called 'metagenomic study' [24]. Screening datasets from metagenomic studies for viral sequences opens up new possibilities for the identification of unknown viruses, especially those which cannot be detected by traditional techniques such as cultivation [20]. Identification of viral sequences in NGS datasets is achieved by comparing the sequences (which may stem from different organisms) obtained from a sample against known sequences in a database using sequence homology search tools such as BLAST [20, 25]. However, the algorithms that are currently employed for sequence comparison only allow the identification of sequences that are closely related to sequences that have already been deposited in databases [20]. Furthermore, NGS technology generally produces a high number of short reads (e.g., most Illumina sequencing platforms yield over 100 million reads, each of between 50 to 600 base pairs in length). This large amount of data makes the

[1] Defined as the *"first temporal emergence of a pathogen in a human population which was related to the increase in distribution, increase in incidence, or increase in virulence, or other factor which led to that pathogen being classified as an emerging disease"* [15].

analysis of such datasets more computationally expensive, in some cases dramatically so [20, 26, 27].

While metagenomic studies have improved our understanding of virus diversity in a wide range of hosts, the following evidence suggests that a large number of viruses still have not been detected. First, conspicuous gaps exist in the known range of viral diversity, such as the almost complete absence of RNA viruses from archaea, and the fact that most known bacteriophages have DNA genomes. Biological explanations for the limited virus diversity in these hosts include the peptidoglycan-containing cell wall in bacteria or the extreme environments inhabited by archaea [28]. However, given the ubiquity of viral types infecting other hosts, these gaps suggest to some that current detection methods may be unable to detect such viruses [29]. Second, based on the assumption that each of the 8.7 million eukaryotic species on earth carries 10 species-specific viruses, Geoghegan *et al.* (2017), suggest that over 99.9% of the existing viral diversity has not yet been detected [30]. Finally, metagenomic studies often yield a high percentage of sequences of unknown provenance [20], sometimes up to 99% [31], some of which may belong to novel viruses.

Since the structure of a protein is more conserved than its sequence [32], being able to detect viral sequences based on structural features may facilitate the identification of highly diverged sequences.

Viruses have fast mutation rates and large population sizes [2]. This makes them a good model system for the study of evolutionary processes, since researchers can observe changes in real time, test hypotheses about virus evolution, such as the effect of point mutations, and test phylogenetic models [2]. Also, since viruses can be altered in the laboratory relatively easily, the effect of mutations can be studied in detail (see for example, Koel *et al.*, 2013 [33]).

However, the ability of viruses to change rapidly can also hinder the study of their evolution, especially when considering evolutionary processes taking place over large timescales: apart from viruses that have integrated into host germ lines, viruses do not leave a fossil record, and the vast majority of viral genomic sequences available today are less than 50 years old. Thus, the study of virus evolution so far has relied on a temporally very limited set of viral samples and sequences [34]. This, together with the fast mutation rates of viruses [35], severely limits our understanding of virus evolution and diversity over time. We do not know what most viruses looked like genotypically even a hundred years ago [34], we know very little about the timescales of virus evolution, as estimates of substitution rates and most recent common ancestor dates often differ between studies (e.g., [36–38]), and we know al-

most nothing about the geographic spread of viruses over time going back more than a few decades.

This PhD thesis is organised into two parts.

Part one describes the application of commonly used sequence matching algorithms to screen for viruses in NGS datasets from remains of individuals that lived thousands of years ago. The viral sequences provide concrete information about the timescales of virus evolution, and the geographic distribution and diversity of the viruses that were recovered. After a brief chapter introducing the field of ancient DNA, the potential of ancient viral sequences for the study of virus evolution, and the datasets studied (Chapter 1), I present the results on three different viruses where ancient sequences could be found: hepatitis B virus (Chapter 2), human parvovirus B19 (Chapter 3), and variola virus (Chapter 4). Chapter 5 discusses these overall findings, their implications, and possible future avenues for research in this field.

Part two describes an algorithm to perform sequence matching based on predicted structural features, in order to identify the origin of highly diverged sequences. The part starts with an overview of the current thinking in the field of sequence comparison and a justification of the approach chosen (Chapter 6). Chapter 7 provides an overview of the matching algorithm. The part concludes with a discussion of the current shortcomings of the algorithm, and future work (Chapter 8). The appendices contain details of the implementation, parametrisation, and evaluation of the algorithm, as well as some preliminary results.

PART I: USING ANCIENT DNA TO STUDY VIRUS EVOLUTION

CHAPTER 1: INTRODUCTION

1.1 AN INTRODUCTION TO THE FIELD OF ANCIENT DNA RESEARCH

Over the last ten years, it has become possible to sequence genetic material from archaeological and environmental samples that are thousands of years old, so-called 'ancient DNA' (aDNA)[1]. The ancient sequence data, combined with data from modern individuals, have revolutionised our understanding of human evolution and past population dynamics [40], giving us more detailed insight into migration, admixture and replacement of human populations over time (e.g., [41–48]), the interbreeding of modern humans with extinct hominins, namely Neanderthals and Denisovans [49, 50], and human adaptation to the environment [51], such as high altitude [52], diet [41, 53, 54], and lifestyles [55]. The field of aDNA research is relatively young, with the first aDNA sequences, from the quagga (an equid species that went extinct in 1883), published in 1984 [56]. Since then, the development of Polymerase Chain Reaction (PCR) in the mid 1980s, modified extraction techniques specifically targeting DNA fragments as short as 35 base pairs, and Next Generation Sequencing (NGS) technology, have improved the efficiency of sequencing of DNA from ancient remains [57]. The first complete ancient human genome and the first draft Neanderthal and Denisovan genomes were all published in 2010 [40, 50, 58, 59]. The first multi-individual study involving ancient humans was published in 2012, presenting data from four approximately 5000 year old individuals from Scandinavia [60]. The oldest aDNA sequences available to date are from Greenlandic ice cores (450–700 thousand years old) [61] and from a Middle Pleistocene horse (560–780 thousand years old) [62].

A challenge when working with DNA recovered from ancient samples is post-mortem fragmentation and chemical modification of the genetic material [63]. While an organism is alive, DNA damage that naturally arises from radiation, metabolic and hydrolytic processes, or exposure to chemicals such as polycyclic aromatic hydrocar-

[1]The term 'ancient DNA' is somewhat loosely used to indicate DNA sequences retrieved from *"museum specimens, archaeological finds, fossil remains, and other unusual sources of DNA"* [39]. However, there is no fixed age from which a sequence or sample is considered 'ancient'. For the purpose of this chapter, I am using 1950 Current Era (CE) as a cut-off.

bons, is constantly being repaired by the DNA damage repair system of the cell [64]. However, after death, DNA damage accumulates, stemming from non-enzymatic processes such as spontaneous hydrolysis and oxidation, and enzymatic processes including nucleases, and digestion by micro-organisms [63, 64]. Four types of damage predominate in aDNA: fragmentation, abasic sites (missing bases), cross-linking of DNA (with another DNA molecule, proteins, or sugars), and miscoding lesions (most commonly changes from cytosine to uracil, increasing in frequency towards the fragment termini) [57, 65–67]. Post-mortem damage and fragmentation affects the preservation of DNA, whose long-term survival is favoured by constant low temperatures [64, 68, 69]. Allentoft *et al.,* (2012) showed that the kinetics of DNA fragmentation follow an exponential decay relationship: 30 base pair fragments are estimated to have a half-life of 158,000 and 500 years, at –5 and +25 degrees Celsius, respectively [68]. Furthermore, humidity negatively affects long-term DNA preservation, due to increased depurination [64, 69].

Another major problem when working with ancient DNA is the high chance of contamination and hence confusion with modern DNA, both from human and non-human sources [63,67]. Especially up to the mid-2000s, aDNA researchers published a number of studies based on false-positive results [67]. Examples include reports of the sequencing of ancient DNA from archaeal and bacterial 16S rDNA in rock salt 11 to 425 million years old [70], from 80 million year old dinosaur bones [71], and also a study describing sequences from a 2,400 year old Egyptian mummy [72]. None of these reports could be verified and all are now considered to involve contamination [63]. Reasons for this conclusion include the following. 1) The half-life predictions in Allentoft *et al.*, (2012) [68] suggest that the survival of sequence fragments of the lengths found in the three studies [70–72] (700 – 1000 base pairs [70], 174 base pairs [71], and 3,400 base pairs [72]), is highly unlikely. 2) The supposed dinosaur sequences [71] were later found to be of mammalian origin, most likely from a human [73, 74]. 3) The close sequence similarity of the archaeal and bacterial 16S rDNA to modern sequences [70] is indicative of modern contamination, given the age of the samples.

The aforementioned problems with contamination resulted in a set of best-practice recommendations for sample handling and laboratory procedures (often referred to as 'authentication criteria') [63, 75, 76]. Their aim is to reduce the chance of contamination, and to give researchers a means to better ensure that recovered sequences are from the ancient sample. The recommendations are: using physically isolated work areas for work prior to the amplification stage; including negative controls for extractions and PCR; ensuring that amplified DNA fragments exhibit appropriate molec-

ular behaviour (PCR amplification strength should be inversely related to the size of the fragment); reproducing results from the same and different extracts from the same sample; verify a number of PCR sequences by cloning of amplified products, to assess contamination, damage, and to detect the presence of nuclear insertions; independently replicate results in a different laboratory; show the preservation of other biomolecules that correlate with DNA survival; and quantify the copy number of the DNA target by Real-Time PCR [63, 75, 76].

The advent of NGS technology has been crucial for the recent advances in research on aDNA [57, 77]. The ability of NGS technology to sequence short DNA fragments, as well as the fact that it alleviates the need to clone PCR products (in contrast to the earlier PCR based techniques and Sanger sequencing), increases the amount of retrievable DNA from ancient samples [57]. When sequencing aDNA, one of two approaches is used [57]. First, whole genome shotgun sequencing, which aims to sequence all genetic material in a sample. Second, capture enrichment, which uses probe sequences designed to bind to specific regions of the genome(s) of interest. Capture methods are used to answer specific questions (by only targeting genomic regions, or whole genomes of interest) and are cheaper than whole genome shotgun sequencing. Whole genome shotgun sequencing can also be used to address specific questions, but in addition makes it possible to analyse the genomes of all organisms present in the sample, including those that were not the focus of the initial investigation. For example, the whole genome shotgun sequencing performed on 101 individuals in Allentoft *et al.*'s (2015) paper on the population genomics of Bronze Age Eurasia [41] made it possible to later screen the datasets for microorganisms. This resulted in the identification of *Yersinia pestis* [78], as well as the hepatitis B virus and human parvovirus B19 sequences presented in this thesis [79, 80].

Widespread adoption of NGS technology for aDNA sequencing resulted in modifications to the original recommendations for ancient DNA authentication outlined above [63, 75, 76], which were developed before the advent of NGS technology [77]. While the recommendations for physically isolated work areas and the inclusion of negative controls remain important, others, such as the appropriate molecular behaviour and cloning of PCR products are no longer relevant [77]. Independent replications in different laboratories are also not done routinely [81]. New recommendations specifically for the use of NGS technologies include the identification of typical aDNA damage patterns on the sequences (in particular the accumulation of cytosine $\rightarrow$ thymine and guanine $\rightarrow$ adenine changes towards the read termini [67]) and, for human sequences, the internal consistency of the data (e.g., haploid human sequences should not show any evidence of polymorphic positions) [77].

Most research on aDNA has been and still is focused on humans. By the end of 2017, aDNA sequences from 711 humans had been published. This is only a small fraction of the ancient individuals that have already been sequenced [82]. For an overview on what has been learned about human evolution using aDNA, see e.g., [40, 82, 83]. Ancient DNA has now also been sequenced from a wide variety of non-human sources, including bone and teeth of horses [62], goats [84], wolves [85], environmental samples [86], and even parchment [87].

1.2 APPLYING ANCIENT DNA TECHNIQUES TO THE STUDY OF MICROORGANISMS

Physicians and scholars have long documented occurrences of disease epidemics affecting humans. A depiction of an individual with one leg shorter than the other and using a walking stick on a 3,500 year old stele in ancient Egypt has been interpreted as representing a polio victim [88]. There are descriptions of a smallpox-like disease, for example in the Vedic text Sushruta Samhita ~100 Before Current Era (BCE), by Ahrun of Alexandria in 622 CE, and by the physician Al-Rhazi in Baghdad in 910 CE [10]. The scholars Thucydides and Galen described the plague of Athens (430 – 426 BCE) and the Antonine plague (165 – 180 CE), respectively. Both of these have been attributed as due to a variety of diseases, including smallpox, measles, bubonic plague, and lassa fever [89]. Although these observations show that humans have long suffered from infectious diseases, descriptions of diseases in early historical records are often ambiguous, as are conclusions drawn based on skeletal pathologies, which can often be caused by multiple pathogens and require long-lasting infections to manifest [90]. Direct evidence of the pathogens involved in past epidemics has mostly been lacking. Given the historical impact of major human epidemics, the uncertainty about their causes, and the potential for future outbreaks due to the same or related pathogens, the application of aDNA techniques to the study of microorganisms, in particular human pathogens, was a logical next step in aDNA research.

As with other research on aDNA performed up to the mid-2000s, the majority of the early attempts to sequence microorganisms from ancient remains did not follow the recommendations for aDNA authentication, or were later judged as probable contaminations [91]. For example, Tsangaras and Greenwood (2012) found that of 43 studies on ancient microorganisms published before 2006, only 15 used physically isolated work areas specifically for the work on aDNA [91]. Similar to studies focusing on the human genome, the advent of NGS technology and development of recommendations

for the sequencing and authentication of aDNA, now allows researchers to establish the authenticity of the ancient sequences with higher confidence [34].

To date, the majority of studies on ancient microorganisms have focused on bacteria, and in particular those causing disease in humans [92, 93]. Unlike in aDNA studies that target the host genome, in which every sample will yield important information (given sufficient DNA preservation), a high number of individuals may need to be screened to detect a pathogen. The number of positive individuals detected will most likely be lower than the number of infected individuals, due to the imperfect diagnostic test, the need for long-term preservation of the pathogen, and whether individuals died during the viraemic or bacteraemic phase of the infection (if any). Thus, a focus has been on bacteria causing chronic and long-term infections (such as *Myobacterium tuberculosis* and *Myobacterium leprae*), since there is a higher possibility that skeletal lesions develop, making it easier to identify samples for sequencing [90, 94]. The first complete sequence of an ancient human pathogen, *Yersinia pestis*, was published in 2011 [95], and further publications on the same bacterium followed [78, 96–98]. The work on ancient *Yersinia pestis* showed that the bacterium has been causing human infections at least since the Neolithic, and that the neolithic and bronze age strains may have been unable to transmit efficiently via the flea vector [78, 92, 96–98]. Since then, ancient sequences for at least twelve human bacterial pathogens have been published [92], including *Helicobacter pylori* (leading to stomach ulcers or gastric carcinoma) [99], the causative agents of tuberculosis (*Myobacterium tuberculosis*) [100], leprosy (*Myobacterium leprae*) [101, 102], brucellosis (*Brucella melitensis*) [103], and syphillis (*Treponema pallidum*) [104], as well as *Salmonella enterica*, which causes enteric fever [105]. In addition, ancient sequences have been recovered from the protozoan parasite *Plasmodium falciparum*, the causative agent of malaria [106]. In addition to studies targeting particular microorganisms, a number of ancient microbiomes have been sequenced from various ancient specimens, including from calcified dental plaque [107] and calcified abscesses from the remains of a woman [108].

1.2.1 Viral sequences isolated from ancient remains

Following the recovery of ancient bacterial sequences from whole genome shotgun sequencing datasets of bronze age individuals [41, 78], an analogous search for evidence for viruses in ancient remains and the recovery of their genomes using similar techniques became a focus. Table 1.1 lists the studies published on ancient viruses up to June 2019. The first successful attempt to extract a viral sequence from ancient

remains was published by Taubenberger *et al.,* in 1997 [109]. The paper presents nine sequence fragments of five segments of the influenza A/H1N1 virus that caused the 1918 influenza pandemic, isolated from formalin-fixed, paraffin-embedded lung tissue samples [109]. The oldest DNA, RNA and human viruses sequenced to date are as follows. The oldest putative viral sequence is from a 140,000 year old tomato mosaic tobamovirus from Greenlandic ice cores, but its authenticity has been questioned based on close sequence similarity between the ancient and the modern viral sequences [110]. The oldest ancient RNA virus sequence is from a 1000 year old zea mays chrysovirus isolated from maize kernels [111]. The oldest human viral sequence is a 7,000 year old hepatitis B virus sequence isolated from a human tooth [112]. Prior to May 2018, the oldest properly authenticated ancient exogenous human viral sequence was from a 450 year old hepatitis B virus isolated from an Italian mummy [113].

In addition to the oldest viral sequences mentioned above, younger sequences have also been recovered, including a 350 year old hepatitis B virus sequence from a Korean mummy [114], two variola virus sequences from 360 and 170 years ago [115, 116], and a partial 300 year old variola virus sequence [117].

Compared to other microorganisms, the study of ancient viruses is complicated by three factors [34]: the first is that many viruses have RNA genomes (the master species list collated by the International Committee on Taxonomy of Viruses lists 2346 out of 5559 virus species with a single- or double-stranded RNA genome). Most aDNA studies target human DNA only, since viruses are not the main focus of the studies, and the resulting sequences contain all the information needed for the study of the human genome. In those studies, RNA is not amplified at all. This makes the discovery of exogenous viruses with an RNA genome from those sample preparations impossible. Second, RNA genomes have been assumed to be less stable over time post-mortem [64] and therefore not as well preserved as their DNA counterparts, possibly again influencing researchers' motivation to study ancient RNA viruses[2]. Finally, the sampling material used for aDNA studies is often taken from teeth or bone (mainly the petrous bone), due to the high endogenous DNA content of those materials, and also because teeth and bones are often the only tissues that are preserved. This has limited the viruses that have been detected to a relatively small set of DNA viruses that cause either chronic or lethal infections with high viral titres in the blood, such as hepatitis B virus, parvoviruses, anelloviruses, orthopoxviruses, adenoviruses, and herpesviruses.

[2]Little is known about the way viral genetic material is preserved in ancient samples. We do not know if the genetic material survives inside capsids or outside, which may affect the preservation of the genetic material [118].

Virus	Authors	Year of publication	Description	Authentication	Reference
Anellovirus	Bédarida *et al.*	2011	One sequence, human dental pulp, 200 years old.	1, 2, 9, 10	[119]
Avipoxvirus	Parker *et al.*	2011	19 sequences, finches and mockingbirds from the Galapagos islands, 120 years old.	1, 2, 3, 4, 6	[120]
Barley stripe mosaic virus	Smith *et al.*	2014	One sequence, isolated from a Barley grain, 750 years old.	1, 2, 8, 9, 10	[121]
Barley yellow dwarf virus	Malmstrom *et al.*	2007	Coat protein sequences, isolated from herbarium specimens of grasses, 1917–1942 CE.	1, 2, 4, 6	[122]
Hepatitis B virus	Bar Gal *et al.*	2012	One sequence, isolated from a Korean mummy, 16th century.	1, 2, 3, 6, 7, 10	[114]
	Patterson Ross *et al.*	2018	One sequence, isolated from an Italian mummy, 16th century.	1, 2, 3, 8, 10	[113]
	Krause-Kyora *et al.*	2018	Three sequences, isolated from teeth, 1000, 5000, and 7000 years old.	1, 2, 3, 8, 9, 10	[112]
	Mühlemann *et al.*	2018	12 sequences, isolated from teeth and bones, 822 to 4488 years old.	1, 2, 3, 8, 9, 10	[79]
Influenza A/H1N1	Taubenberger *et al.*	1997	One sequence, isolated from a formalin-fixed, paraffin-embedded, lung tissue sample, 100 years old.	3, 4, 7, 9, 10	[109]
Monkeypox	Tiee *et al.*	2018	93 sequences, isolated from museum specimens of African striped squirrels, up to 130 years old.	1, 2, 4, 10	[123]
Papillomavirus	Fornaciari *et al.*	2003	One sequence, isolated from an Italian mummy, 16th century.	2, 4, 7	[124]
	Larsen *et al.*	2018	One sequence, isolated from packrat middens, 27,000 years old.	1, 2, 10	[125]
Human Parvovirus B19	Toppinen *et al.*	2015	43 sequences, isolated from putative Finnish casualties during WW II, approx. 75 years old.	1, 2, 9, 10	[126]
	Mühlemann *et al.*	2018	10 sequences, isolated from teeth and bones, 500 to 6800 years old.	1, 2, 3, 8, 9, 10	[80]
Pithovirus	Legendre *et al.*	2014	One sequence isolated from a sample of Siberian permafrost which was grown in culture, dated to 30,000 years old.	11	[127]
Simian T-cell leukemia virus	Calvignac *et al.*	2008	One sequence, isolated from an African green monkey, 1913.	1, 2, 3, 4, 5, 6, 7, 10	[128]
Human T-cell leukemia virus	Li *et al.*	1999	One sequence, isolated from an Andean mummy, 1500 years old.	2, 3, 4, 7, 10	[129]
Tomato mosaic tobamovirus	Castello *et al.*	1999	17 sequences of tomato mosaic tobamovirus, isolated from ice cores from Greenland, 500–140,000 years old.	2, 4, 6	[130]
Variola virus	Biagini *et al.*	2012	One sequence fragment, isolated from a mummy in Siberian permafrost, 17th to early 18th century.	1, 2, 6	[117]
	Duggan *et al.*	2016	One sequence, isolated from a Lithuanian mummy, 17th century.	1, 2, 3, 8, 9, 10	[116]
	Pajer *et al.*	2017	Two sequences isolated from Czech museum specimens, 60 and 160 years old.	2	[115]
	Mühlemann *et al.*	unpublished	Three high- and ten low-coverage sequences, isolated from teeth and bones, 550 – 1900 CE.	1, 2, 3, 8, 9, 10	unpublished
Zea mays chrysovirus 1	Peyambari *et al.*	2018	three sequences of Zea mays chrysovirus 1, isolated from maize kernels from Arizona, US, 1,000 years old.	1, 2	[111]
Various	Appelt *et al.*	2014	Metagenomic characterisation of a 14th century coprolite.	1, 2	[131]
	Ng *et al.*	2014	Caribou feces associated virus and Northwest Territories cripavirus, isolated from frozen caribou feces, 700 years old.	1, 2, 6, 7	[132]

Table 1.1: Previously published studies of ancient exogenous viruses. Only studies with sequences from before 1950 were included. For a review of more modern (as well as ancient) virus studies, see Tsangaras and Greenwood, 2018 [34]. Rows are sorted alphabetically by virus name and then by year of publication. The 'Authenticated' column uses the following abbreviations: 1: Physically isolated work areas. 2: Negative controls (extraction, PCR controls). 3: Repeated amplification from the same and different extracts. 4: Inverse correlation between amplification efficiency and length of amplification. 5: Associated remains. 6: Reproduction in a second independent laboratory. 7: Cloning of amplification products and sequencing of multiple clones. 8: Damage patterns. 9: Phylogenetic analyses support ancient origin. 10: Preservation of host DNA. 11: The virus was recovered using tissue culture techniques.

1.2.1.1 How ancient viral sequences could improve our understanding of virus evolution

Understanding how viruses change over time or in response to the environment is relevant for human and animal health. For example, it can help to understand the development of antiviral resistance and immune evasion, as well as improve our general knowledge of evolutionary processes, such as extinction, substitution rates, and recombination. Viruses have fast mutation rates, large population sizes, and short generation times [2]. This allows researchers to observe evolutionary processes in real time, which is rarely possible in higher organisms. A good example is research on the influenza virus. Researchers have been able to visualise the antigenic evolution of seasonal influenza A/H3N2 [133], understand the global circulation patterns of the virus [134], and can identify specific mutations causing antigenic changes that result in the escape from prevailing immunity [33]. This work has informed vaccine choices for seasonal influenza. It has also added to the general understanding of how immunity due to influenza infection and vaccination develops over time [135].

Viruses, like any organism, evolve via natural selection. Diversity is a prerequisite for adaptive change in the face of selection pressure. In viruses, diversity arises as a result of processes including point mutations, recombination, re-assortment (in viruses with segmented genomes), and horizontal gene transfer [2]. How quickly viruses accumulate changes in their genome is a central part of our understanding of their natural history, including pathogenesis [136], the likelihood of the emergence of novel viruses [137], and the timescales of virus evolution [35]. The rate at which mutations occur in the genome per unit of time, round of replication, or generation is called the 'mutation rate'. Viral mutation rates approximately correspond to the fidelity of the polymerase used in replication: viruses with an RNA genome, which use a more error prone RNA-dependent RNA polymerase for replication, mutate rapidly (on the order of 10^{-5} – 10^{-3} substitutions per site per cell infection), while DNA viruses, replicating using a DNA polymerase with the ability to correct errors, mutate more slowly (on the order of 10^{-8} – 10^{-5} substitutions per site per cell infection). The mutation rate of retroviruses, which replicate using reverse transcriptases, falls between the two [35, 138]. Mutation rates can be estimated by comparing the changes in the genome sequences of the original virus and the progeny virus after one round of infection. However, this method can be confounded by sequencing error or effects of selection. An alternative is to use Luria–Delbrück fluctuation tests, which involve counting rare mutations to an easily observable phenotype, such as resistance to a drug [139]. This frequency is then adjusted to account for the number of generations and genome replications to obtain the mutation rate [35].

While many of the mutations that arise are deleterious or neutral, some may be selected for, due to their beneficial phenotype, and ultimately reach high frequency in the virus population. Possible phenotypic changes include adaptation to different cell types or receptors, immune evasion, or replication in different hosts (e.g., [33, 140, 141]). The rate at which mutations are fixed at the population level per site in the genome per unit of time is called the 'substitution rate'. Four factors govern the substitution rate: the mutation rate, generation time, effective population size, and the fitness effects of mutations [35]. In order to infer substitution rates and calculate divergence dates between sequences, we need a model of how mutations accumulate over time, a 'molecular clock model'. The original molecular clock model suggested that mutations accumulate at the same rate on all lineages in a phylogeny [142, 143]. Since many datasets violate the assumption of the original molecular clock [144], relaxed molecular clock models were developed, in which rates may vary between branches [142, 145]. There are different ways to vary the rates between branches of a phylogenetic tree, such as by assuming 1) a small number of different rates, clustered across the tree (local clock); 2) a small number of different rates, but not clustered across the tree (discrete clock); 3) different rates for all branches, which either correlate between branches (autocorrelated relaxed clock), or are independently and identically distributed (uncorrelated relaxed clock) [145]. The molecular clock can only give measurements on an absolute timescale (such as days or years) if it is calibrated with some form of time information [142]. Calibrations can be of three types:

- Calibration of nodes in the tree using external information. Calibrations can be based on the fossil record, divergence times of the host based on the assumption of co-evolution, or on biogeographic processes such as island formation [142, 146, 147].
- The substitution rate can be known from previous work. For example, if the virus being studied is integrated in the germ line of the host, the host substitution rate can be used.
- If the virus population is measurably evolving[3], sequences sampled at different time points, so-called heterochronously sampled sequences, can be used to calibrate the molecular clock [142, 148, 149]. Datasets of heterochronously sampled sequences that span a wide time range are rare, since most available viral sequences are less than 50 years old.

[3] A measurably evolving population is defined as *"a population from which molecular sequences can be taken at different points in time, among which there are a statistically significant number of genetic differences."* [148]. This requires either a wide range of sampling times, or a high mutation rate [148].

Methods for the inference of substitution rates can be based on linear regressions of root-to-tip distances in a phylogenetic tree and sampling dates, and on maximum likelihood or Bayesian frameworks [35, 142]. Probably the most commonly used software package for the inference of substitution rates for viruses is that provided by BEAST [150, 151]. BEAST is based on Bayesian inference and uses the Markov chain Monte Carlo algorithm to explore different model parameter values. This allows the simultaneous inference of the substitution rate, tree topology, and demographic processes. Methods used to infer substitution rates are based on two well-known assumptions, both of which may lead to an artificial inflation of the rate, if violated [35]. Firstly, they assume that the sampled sequences are not recombinant, and secondly that the sequences contain only mutations that have been fixed. Since the latter is rarely the case, the substitution rate that is inferred is thus likely a composite of the mutation and the substitution rates.

On the basis of the fidelity of different polymerases, it has been proposed that RNA viruses should have faster substitution rates than DNA viruses [35]. In general, RNA viruses have been found to be rapidly-evolving, with substitution rates in the range of 10^{-5} to 10^{-2} substitutions per site per year (s/s/y) and DNA viruses have been considered slowly-evolving, with rates between 10^{-9} to 10^{-6} s/s/y [2, 152, 153]. However, there are exceptions to the general trend of rapidly-evolving RNA and slowly-evolving DNA viruses. For example, some RNA viruses evolve with a rate that is slower than that which was generally assumed for RNA viruses, including the simian foamy viruses ($1.7\text{x}10^{-8}$ s/s/y) [154] and rodent associated hantavirus ($0.7\text{x}10^{-7}$ – $2.2\text{x}10^{-6}$ s/s/y) [155]. Also, some single-stranded DNA viruses are thought to evolve at rates closer to those of RNA viruses, for example canine parvovirus ($1.7\text{x}10^{-4}$ s/s/y in the capsid protein VP2 and $7.9\text{x}10^{-5}$ s/s/y in the nonstructural protein NS1) [156], human parvovirus B19 (1.0 – $4.0\text{x}10^{-4}$) [126, 157, 158], circovirus SEN-V ($7.32\text{x}10^{-4}$ s/s/y) [159], and tomato yellow leaf curl virus ($4.63\text{x}10^{-4}$ s/s/y) [160]. While this list is not exhaustive, it is interesting to note that the substitution rates for the slowly-evolving RNA viruses have all been inferred using external calibrations, while the substitution rates for the rapidly-evolving DNA viruses have all been inferred using heterochronously sampled sequences. The trend that rates estimated over a longer timescale (using calibrations from fossils, assumptions about virus-host co-divergence, or other external events) are significantly lower than rates estimated over shorter timescales (using modern heterochronously sampled sequences), is known as the 'time dependent rate phenomenon' [38, 152, 161, 162]. The time dependent rate phenomenon has been observed in many viruses, including hepatitis B virus and JC Polyomavirus. For hepatitis B virus, Zhou and Holmes (2007), inferred a rate of $7.72\text{x}10^{-4}$ s/s/y using heterochronously sampled sequences [37] whereas us-

ing external calibrations based on human migrations, Paraskevis *et al.* (2015) found the rate to be 3.7×10^{-6} s/s/y [36]. For JC polyomavirus, under the assumption of co-divergence with humans as they left Africa, a rate of approximately 4×10^{-7} s/s/y was inferred [163], but using sequences sampled over the span of 34 years, the rate was found to be two orders of magnitude higher [164]. The causes of the time dependent rate phenomenon are debated, and include that substitution rates observed over short time scales are inflated due to the presence of transient mutations, sequencing error, changes in virus biology and selection pressure over time, calibration errors (such as assuming that the time of population divergence and genetic divergence correspond), and substitution model inaccuracy resulting in underestimation of saturation [161].

Before the advent of aDNA sequencing techniques, virologists assumed that viruses do not leave a fossil record [2], with viruses that have integrated into the germ line of the host genome [165], and the sequences from before 2018 listed in Table 1.1, being the exceptions. As a result of the sparsity of ancient viral sequences, we only have a limited understanding of historical viral diversity, even going back just a few decades. It is unclear if viruses in the past were closely related to their modern counterparts, or whether they were much more diverged, as might be expected based on their high mutation rates. Furthermore, even though phylogenetic tools such as BEAST have been widely used to study virus evolution and infer substitution rates and most recent common ancestor dates, it has been impossible to test them extensively on non-artificial sequences older than a few decades. In terms of public health and pandemic preparedness, it would be informative to have an understanding of the variation of a particular virus that existed in the past, in order to learn of types of changes that may arise in the future.

Nevertheless, the small number of viral sequences a few hundred years old (Table 1.1) and sequences from endogenous viral elements (EVEs) have provided valuable information about the evolutionary history of viruses. This is illustrated by research on the influenza A/H1N1 virus that caused the 1918 influenza pandemic.

In 1997, Taubenberger *et al.* presented partial sequences of five segments of the A/H1N1 virus which caused the 1918 influenza pandemic, and complete sequences for all eight segments followed. Initially, Taubenberger *et al.* (1997) found the virus to be most closely related to strains that infect humans and swine [109]. Later work pointed to the emergence of the virus after a human H1 virus acquired the neuraminidase and internal protein genes from an avian virus [166]. Access to sequences from the influenza 1918 virus also allows researchers to study mutations that may lead to human adaptation of the virus. Hemagglutinin sequences recovered from

13 individuals that died at different times during the pandemic suggest an adaptation from more avian-like to more human-like receptor binding over the course of the pandemic, due to a change from glycine to aspartic acid at position 225 (H3 numbering) [167]. Tumpey *et al.* (2005) reconstructed the influenza 1918 virus using reverse genetics and found that it caused more severe lung pathology and reached 125 to 39,000 times higher virus titres in the lungs of mice four days post infection than the epidemic H1N1 viruses circulating between 1930 to 1990 that were used as controls [168]. The influenza 1918 virus caused death of embryonated chicken eggs in contrast to the control epidemic H1N1 viruses and had the ability to replicate in the absence of trypsin [168]. It also lacked a multibasic cleavage site in the hemagglutinin, which is found in highly pathogenic influenza H5 and H7 viruses [109]. However, although the 1918 influenza virus was more virulent than epidemic H1N1 viruses, the additional virulence is not sufficient to account for the high mortality rates that were observed during the pandemic [169]. An increased expression of pro-inflammatory chemokines and suppression of the type I interferon response, together with the exposure history to previously circulating influenza viruses, may have contributed to the high death rate during the pandemic [169–171].

While the influenza virus causing the 1918 pandemic is an example of a relatively recent sequence, important information can be gained from sequences from EVEs that can be millions of years old. EVEs are sequences of viruses that have integrated into the germline of the host and are passed on from generation to generation [172]. In animals, most EVEs are derived from retroviruses [165, 173]. However, EVEs have also been found from non-retroviral families, including *parvoviridae, hepadnaviridae, bornaviridae*, and *orthomyxoviridae* [165]. Sequences of EVEs accumulate changes at the rate of evolution of the host, thereby preserving sequence information about the virus that would otherwise most likely be lost due to the fast viral mutation rates [2]. Together with approximate dates of when the integration event took place, researchers can use the sequences of EVEs to calibrate molecular clocks and gain an understanding of how long a virus may have been associated with a certain host. For example, using an endogenous hepadnavirus found in the genome of the zebra finch, Gilbert and Feschotte (2010) found that the integration event took place at least 19 million years ago and were able to infer a substitution rate in the order of 10^{-8} s/s/y, 1000-fold lower than the rate inferred from extant hepadnaviruses [37, 174].

The work done on the 1918 influenza virus and the EVEs indicate that ancient viral sequences can improve our understanding of virus evolution, in terms of studying substitution rates, and understanding virus origins and mutations associated with particular phenotypes. Ancient DNA sequencing techniques have made it possible to

recover ancient viral sequences that are older than the ancient exogenous (i.e., non-integrated) viral sequences that have been previously recovered (Table 1.1). Such viral sequences will allow us to obtain a better understanding about viral substitution rates over longer time scales, as well as past viral diversity and geographic distribution.

1.3 APPROACH AND DATASETS

The work presented in chapters 2, 3, and 4 is based on the screening of aDNA whole genome shotgun datasets comprising of 1653 humans (see Table 1.2). Figure 1.1 shows the locations of the samples, and Table 1.2 provides further descriptions and references. The individuals were sequenced for the purpose of analysing the human genome (see corresponding references in Table 1.2), and some have also subsequently been screened for bacterial pathogens, resulting in the finding of Bronze Age *Yersinia pestis* sequences [78]. The goal of the virus screening was not to provide a characterisation of the entire metagenome or virome of the ancient individuals, but rather to – after the initial screening – identify and describe sequences of specific viruses. The datasets were assembled and sequenced by Prof. Eske Willerslev, Dr. Lasse Vinner, Dr. Ashot Margaryan, Dr. Peter de Barros Damgaard, Prof. Morten Allentoft, Dr. Hugh McColl, Dr. Constanza de la Fuente Castro, Prof. Andrea Manica, Dr. Eppie Jones, and Prof. Turi King. Targeted virus capture was performed by Dr. Lasse Vinner.

Dataset	ENA accession	Number of Individuals	Age range (years)	Reference	Chapter
Viking	unpublished	555	880–1,100	[175]	3, 4
South East Asia	PRJEB26721	32	300–4,400	[176]	3, 4
Iron Age	PRJEB20658	142	500–4,500	[44]	2, 3, 4
Bronze Age	PRJEB9021	168	1,400–5,400	[41]	2, 3, 4
Baikal Hunter Gatherers	PRJEB26349	31	3,800–7,100	[43]	3, 4
Kolyma	PRJEB29700, PRJEB26336	25	800–31,600	[177]	3, 4
Neolithic	unpublished	659	500–7,000	unpublished	3, 4
Various	PRJEB11364, PRJEB18067, PRJEB20614, PRJEB20616, PRJEB13189, PRJEB11995, PRJEB14817, SRP039766	41	1,736–13,770	[178–185]	3, 4

Table 1.2: Human shotgun NGS datasets screened for viruses in this thesis. The 'Chapter' column indicates the chapter(s) of this thesis in which the dataset is relevant.

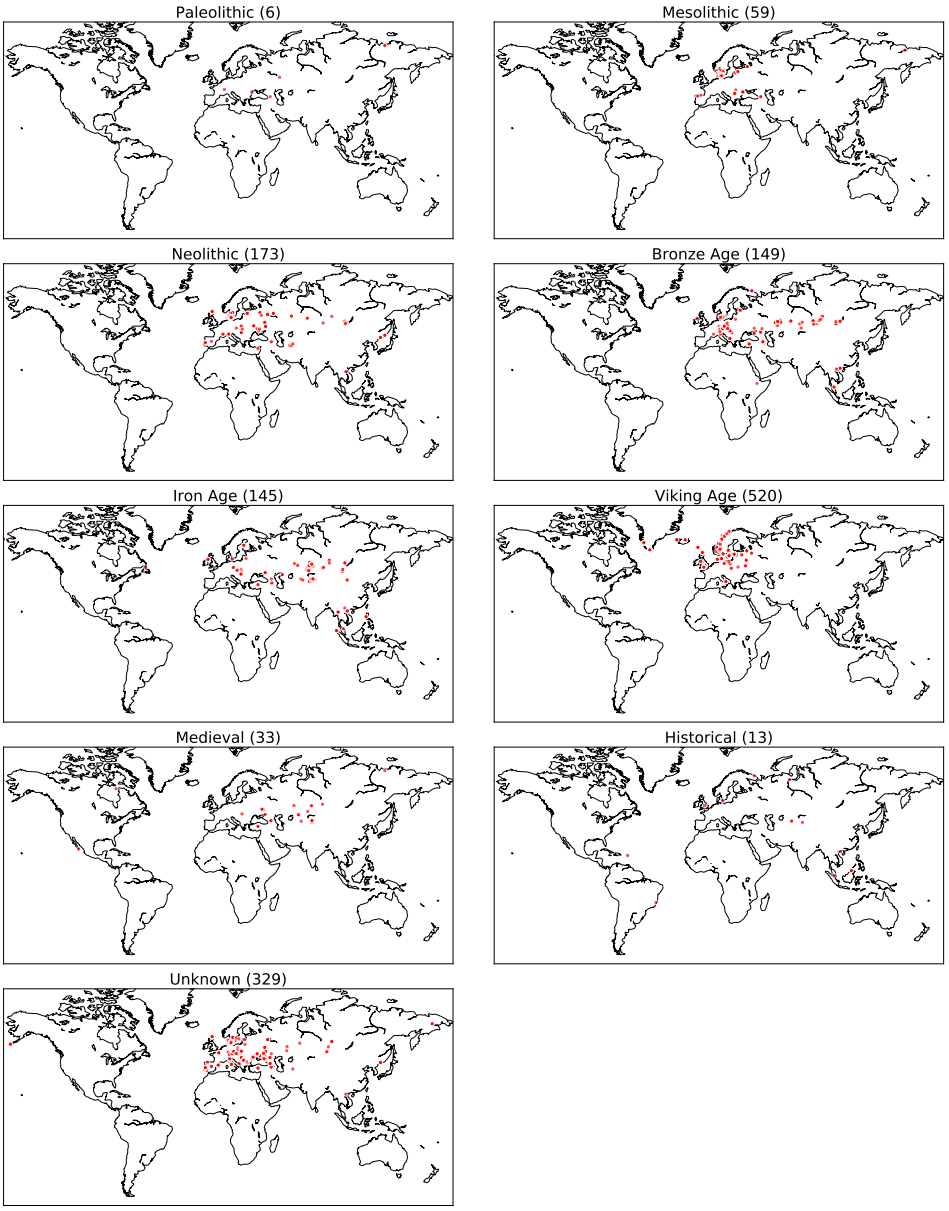


Figure 1.1: Map of the locations of the samples analysed during this PhD. Samples are plotted on different maps according to approximate sample age. Paleolithic: older than 30,000 years, Mesolithic: 30,000 – 10,000 years, Neolithic: 10,000 – 6000 years, Bronze Age: 5,000 – 3,000 years, Iron Age: 3,000 – 1,200 years, Viking Age: 1,200 – 900 years, Medieval: 900 – 300 years, Historical: 300 – present. Location and approximate age information was available for 1098 out of 1653 samples. For the 329 samples in the 'Unknown' panel, the age is not known, but their location is.

CHAPTER 2: ANCIENT HEPATITIS B VIRUSES FROM THE BRONZE AGE TO THE MEDIEVAL PERIOD

PREFACE

A version of this chapter was previously published as (* denotes equal contribution):

Barbara Mühlemann*, Terry C. Jones*, Peter de Barros Damgaard*, Morten E. Allentoft*, Irina Shevnina, Andrey Logvin, Emma Usmanova, Irina P. Panyushkina, Bazartseren Boldgiv, Tsevel Bazartseren, Kadicha Tashbaeva, Victor Merz, Nina Lau, Vàclav Smrčka, Dmitry Voyakin, Egor Kitov, Andrey Epimakhov, Dalia Pokutta, Magdolna Vicze, T. Douglas Price, Vyacheslav Moiseyev, Anders J. Hansen, Ludovic Orlando, Simon Rasmussen, Martin Sikora, Lasse Vinner, Albert D. M. E. Osterhaus, Derek J. Smith, Dieter Glebe, Ron A. M. Fouchier, Christian Drosten, Karl-Göran Sjögren, Kristian Kristiansen, Eske Willerslev. *Ancient Hepatitis B viruses from the Bronze Age to the Medieval period.* Nature, 557, 418–423 (2018).

It has been modified to fit the style of a dissertation.

I performed the screening of the datasets in collaboration with Terry Jones, assembled the consensus sequences, did the phylogenetic analyses (with input by Christian Drosten), authentication via damage patterns, genome properties, figures and tables (except the recombination, and sequence identity figures and tables), and wrote the text in collaboration with Terry Jones and with input from all co-authors. The recombination analysis, genotyping, the analysis of similarity to modern sequences, and authentication via BLASTn was done by Terry Jones. The sequencing work was performed by Peter de Barros Damgaard and Morten Allentoft. PCR confirmation and targeted virus capture of a subset of samples was performed by Lasse Vinner, and capture probes were designed by Anders Hansen. In addition to this description, I have noted in the legends of figures and tables if they were contributed by others.

2.1 ABSTRACT

Hepatitis B virus (HBV) is a major cause of human hepatitis. There is considerable uncertainty about the timescale of evolution of the virus and its association with humans. Here we present 12 full or partial ancient HBV genomes that are between approximately 0.8 and 4.5 thousand years old. The ancient sequences group either within or in a sister relationship with extant human or other ape HBV clades. Generally, the genome properties follow those of modern HBV. The root of the HBV tree is projected to between 8.6 and 20.9 thousand years ago, and we estimate a substitution rate of 8.04×10^{-6} – 1.51×10^{-5} nucleotide substitutions per site per year. In several cases, the geographical locations of the ancient genotypes do not match present-day distributions. Genotypes that today are typical of Africa and Asia, and a subgenotype from India, are shown to have an early Eurasian presence. The geographical and temporal patterns that we observe in ancient and modern HBV genotypes are compatible with well-documented human migrations during the Bronze and Iron Ages [41, 44]. We provide evidence for the creation of HBV genotype A via recombination, and for a long-term association of modern HBV genotypes with humans, including the discovery of a human genotype that is now extinct. These data expose a complexity of HBV evolution that is not evident when considering modern sequences alone.

2.2 INTRODUCTION

HBV is transmitted perinatally or horizontally via blood or genital fluids [186]. The estimated global HBsAg seroprevalence is 3.6%, ranging from 0.01% (UK) to 22.38% (South Sudan) [187]. In high endemicity areas, in which HBsAg seroprevalence is over 8%, 70–90% of the adult population show evidence of past or present infection measured by the presence of any of HBsAg, anti-HBs, or anti-HBc [188–190]. The young and the immunocompromised are most likely to develop chronic HBV infection, which can result in high viraemia over years to decades [186]. Approximately 257 million people are chronically infected and around 887,000 died in 2015 owing to associated complications [190]. Despite the prevalence and public health impact of HBV, its origin and evolution remain unclear [191, 192]. Inference of HBV nucleotide substitution rates is complicated by the fact that the virus genome consists of four overlapping open reading frames [193], and that mutation rates differ between phases of chronic infection [194]. Studies based on heterochronous sequences, sampled over a relatively short time period, find higher substitution rates, whereas rates estimated using external calibrations tend to be lower, leading to a wide range of estimated HBV substitution rates ($7.72\text{x}10^{-4}$ – $3.7\text{x}10^{-6}$ substitutions per site per year) [36, 37, 195]. Human HBV is classified into at least nine genotypes (A–I) based on sequence similarity of at least 92.5% within genotypes [196], with a heterogeneous global distribution [192, 193] (Fig. 2.1a). Attempts to explain the origin of genotypes using human migrations have been inconclusive. The hypothesis that HBV co-evolved with modern humans as they left Africa 60–100 thousand years ago (ka) has been contested owing to the basal phylogenetic position of genotypes F and H, which are found exclusively in the Americas [191]. HBV also infects non-human primates (NHP), and the human and other great ape HBVs are interspersed in the phylogenetic tree, possibly due to cross-species transmission [197]. Given the variability of estimated substitution rates, the incongruence of the tree topology with some human migrations and the mixed topology of the NHP and human HBV sequences in the phylogenetic tree, there remains considerable uncertainty about the evolutionary history of HBV. Recent advances in the sequencing of ancient DNA (aDNA) have yielded important insights into human evolution, past population dynamics [40] and diseases [78, 96]. However, ancient sequences have been recovered for only a handful of exogenous human viruses, including influenza virus (sample approximately 100 years old) [198], variola virus (sample approximately 360 years old) [116] and HBV (samples approximately 340 and 450 years old) [113, 114]. The knowledge gained from these cases emphasizes the general importance of ancient sequences for the direct study of long-term viral evolution. HBV has several characteristics that

make it a good candidate for detection in an aDNA virus study: its long-lasting viraemia during chronicity, with high titres of serum HBV DNA (>2,000 International Units (IU)/mL, 1IU=5 viruses) [2, 186], the relative stability of its virion [199], and its small, circular and partially double-stranded DNA genome [193]. Here we present 12 ancient HBV sequences between 0.8 and 4.5 thousand years old (kya), identified by screening shotgun sequencing datasets from 167 Bronze Age and 137 predominantly Iron Age individuals from central to western Eurasia with a sample age range of approximately 7.1–0.2 kya.

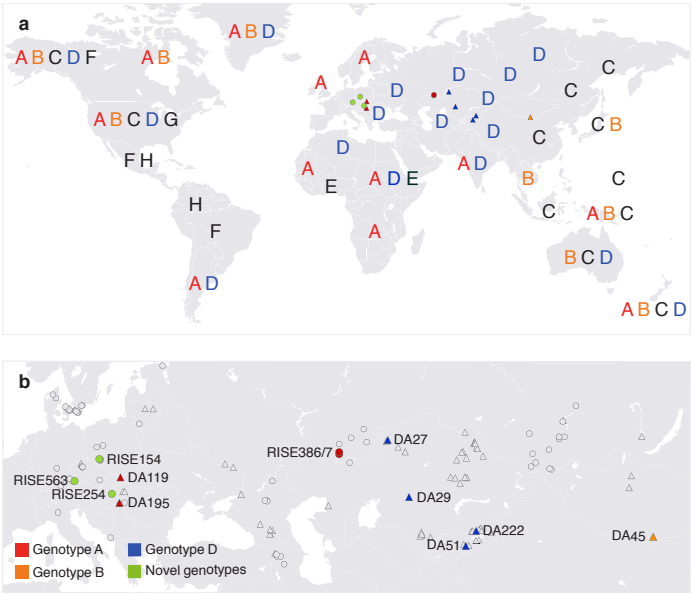


Figure 2.1: Geographical distribution of analysed samples and modern genotypes. a) Distribution of modern human HBV genotypes [192]. Genotypes relevant to this chapter are shown in colour, the others in black. Coloured shapes indicate the locations of the HBV-positive samples included for further analysis. **b)** Locations of analysed Bronze Age samples [41] are shown as circles and Iron Age and later samples [44] are shown as triangles. Coloured markers indicate HBV-positive samples. Ancient genotype A samples are found in regions in which genotype D predominates today, and HBV-DA27 is of subgenotype D5 which today is found almost exclusively in India.

2.3 METHODS

2.3.1 HBV datasets

The following HBV datasets were used in the present study:

Dataset 1
Dataset 1 comprises 26 HBV genomes, covering all species in the *Orthohepadnaviridae*. This includes one sequence each from the human HBV genotypes (A–I, and J), orangutan, chimpanzee, gorilla, gibbon, woolly monkey, woodchuck, ground squirrel, Arctic ground squirrel and horseshoe bat, four sequences from roundleaf bats, and three sequences from tent-making bats, largely following a previous publication [200].

10 sequences from human genotypes (genotype in parentheses): AM282986 (A), D23678 (B), AB117758 (C), AB126581 (D), AB205192 (E), X69798 (F), AB056513 (G), AY090454 (H), AF241409 (I), AB486012 (J).

Four sequences from NHPs: AF305327 (Chimpanzee), AY781180 (Gibbon), AJ131-567 (Gorilla), AF193863 (Orangutan).

12 from other *Orthohepadnaviridae* representatives: KC790373 (Roundleaf bat), KC790374 (Roundleaf bat), KC790376 (Roundleaf bat), KC790375 (Roundleaf bat), KC790377 (Horseshoe bat), KC790378 (Tent-making bat), KC790380 (Tent-making bat), KC790381 (Tent-making bat), NC_004107 (Woodchuck), NC_001484 (Ground squirrel), U29144 (Arctic ground squirrel), AF046996 (Woolly monkey).

Dataset 2
Dataset 2 comprises 124 HBV genomes, from humans and NHPs. This set contains 92 sequences from a previous publication [36] (excluding incomplete sequences), 7 additional genotype D sequences, the Korean mummy genotype C sequence [114], the 12 ancient sequences from the present study and 12 full genomes selected from a set of 9,066 full HBV genomes downloaded from NCBI [201] on 24 August 2017 (Entrez query: `hepatitis b virus[organism] not rna[title] not clone[title] not clonal[title] not patent[title] not recombinant[title] not recombination[title] and 3000:4000[sequence length]`) corresponding to the closest, non-artificial match for each of the ancient sequences. Dates for these sequences were acquired by looking for a date of sample collection in the NCBI entry, or the paper in which the sequence was first published. If a range of dates was mentioned, the mean was used. If no date of sample collection was found in this way, either the year of the publication of the

paper, or the year of addition of the sequence to GenBank was used, whichever was earlier.

85 human sequences from Paraskevis *et al.* (2015) [36] ((sub)genotype in parentheses): AB076679 (A1), AB116084 (A1), AB453988 (A1), AY738142 (A2), GQ477499 (A2), AY934764 (A3), FJ692556 (A3), FJ692598 (A3), FJ692611 (A3), GQ161813 (A3), GQ331046 (A4), AB073858 (B1), AB033555 (B3), AB219429 (B3), AB219430 (B3), AP011089 (B3), AB073835 (B4), AB287316 (B5), AB287318 (B5), AB287320 (B5), AB287321 (B5), DQ463789 (B5), DQ463792 (B5), AB241117 (B), DQ993686 (B), AB111946 (C1), AB112066 (C1), AB112472 (C1), DQ089767 (C1), X75656 (C3), X75665 (C3), AB048704 (C4), AB048705 (C4), AF241411 (C5), AP011100 (C5), AP011102 (C6), AP011103 (C6), AP011106 (C8), AP011108 (C9), FJ899792 (D1), JN642140 (D1), GQ477453 (D2), GQ477455 (D2), JN642160 (D2), JN642163 (D2), JN688710 (D3), JN688711 (D3), GQ922005 (D4), HE974378 (D4), KJ470893 (D4), KJ470896 (D4), KJ470898 (D4), FJ904430 (D7), FJ904436 (D7), AB033559 (D), AB048701 (D), AB048702 (D), AB188243 (D), AB210818 (D), AM494716 (D), AY796031 (D), AY902768 (D), DQ315779 (D), X80925 (D), X75657 (E), X75664 (E), AY090458 (F1a), AB116654 (F1b), FJ657525 (F1b), AY090455 (F2a), AY311369 (F2a), DQ899144 (F2b), DQ899146 (F2b), AB116549 (F3), X75663 (F3), AF223962 (F4), AB166850 (F), AB056513 (G), AB064312 (G), AF405706 (G), AB059660 (H), AB375163 (H), AY090454 (H), AY090457 (H), AB486012 (J).

7 NHP sequences from Paraskevis *et al.,* (2015) [36]: AY330911 (Chimpanzee), AJ131571 (Gibbon), AY781180 (Gibbon), U46935 (Gibbon), AJ131567 (Gorilla), AF193863, EU155824 (Orangutan).

7 additional human genotype D sequences: AB033558, GQ205382, GQ205389, GQ205384, GQ205377, GQ205385, GQ205378.

1 Korean mummy genotype C sequence from [114]: JN315779

12 ancient sequences from the present study: ERS2295383 (DA27), ERS2295384 (DA29), ERS2295385 (DA45), ERS2295386 (DA51), ERS2295387 (DA119), ERS2295388 (DA195), ERS2295389 (DA222), ERS2295390 (RISE154), ERS2295391 (RISE254), ERS2295392 (RISE386), ERS2295393 (RISE387), ERS2295394 (RISE563).

The 12 modern sequences that are closest to each of the ancient sequences: KP322603 (DA27), KC875319 (DA29), KP341007 (DA45), KP322600 (DA51), FN545831 (DA119), EU859952 (DA195), KP322602 (DA222), FM209516 (RISE154), AF222323 (RISE254), GQ331047 (RISE386), KC774243 (RISE387), AB032433 (RISE563).

Dataset 3

Dataset 3 comprises 124 HBV genomes, from humans, NHPs and a variety of other *Orthohepadnaviridae* host species, including woolly monkey, roundleaf bat, tent-making bat, ground squirrel, Arctic ground squirrel, woodchuck and snow goose. This set contains 113 sequences that were obtained from the union of 91 sequences from Paraskevis *et al.,* (2015) [36] and 29 from Drexler *et al.,* (2013) [200], plus 11 additional sequences (giving 124 sequences in total).

113 sequences formed by the union of:

91 sequences from Paraskevis *et al.,* (2015) [36]. These are the 92 listed above in Dataset 2, excluding AF405706, which resulted in a duplicate once gaps were removed from the multiple sequence alignment including all ancient genomes. 27 sequences from Drexler *et al.,* (2013) [200]: AB056513, AB117758, AB126581, AB205192, AB486012, AF046996, AF111000, AF193863, AF241409, AF305327, AJ131567, AJ131568, AM282986, AY090454, D23678, KC790373, KC790374, KC790375, KC790376, KC790377, KC790378, KC790380, KC790381, NC_001484, NC_004107, U29144, X69798.

11 additional sequences: AB032433, AB493845, DQ298161, DQ336682, GQ205385, GQ331047, GQ475343, GQ922004, JN040772, KF679991, KJ470892.

Dataset 4

Dataset 4 comprises 3,505 HBV genomes. Of these, 3,384 are from a previous publication [202], divided into ten human genotypes. To these, we added 17 chimpanzee, 56 gorilla, 12 gibbon and 36 orangutan full HBV genome sequences downloaded from NCBI on 18 January 2017, resulting in 14 genome categories.

3384 HBV genomes from Bell *et al.,* (2016) [202].

17 Chimpanzee sequences: AB368296, AF222322, AF222323, AF305327, AJ131575, D00220, FJ798098, FJ798099, JQ664502, JQ664503, JQ664504, JQ664505, JQ664506, JQ664507, JQ664508, JQ664509, U46935.

56 Gibbon sequences: AJ131568, AJ131569, AJ131570, AJ131571, AJ131572, AJ131573, AJ131574, AY077735, AY077736, AY781177, AY781178, AY781179, AY781180, AY781181, AY781182, AY781183, AY781184, AY781185, AY781186, AY781187, EU155821, EU155822, EU155823, EU155824, EU155825, EU155826, EU155827, EU155828, EU155829, HQ603058, HQ603059, HQ603060, HQ603061, HQ603062, HQ603063, HQ603064, HQ603065, HQ603066, HQ603067, HQ603068, HQ603069, HQ603070, HQ603071, HQ603072, HQ603073, HQ603074, HQ603075,

HQ603076, HQ603077, HQ603078, HQ603079, HQ603080, HQ603081, HQ603082, KT893897, U46935.

12 Gorilla sequences: AJ131567, FJ798095, FJ798096, FJ798097, JQ664502, JQ664503, JQ664504, JQ664505, JQ664506, JQ664507, JQ664508, JQ664509.

36 Orangutan sequences: AF193863, AF193864, EU155821, EU155822, EU155823, EU155824, EU155825, EU155826, EU155827, EU155828, EU155829, HQ603058, HQ603059, HQ603060, HQ603061, HQ603062, HQ603063, HQ603064, HQ603065, HQ603066, HQ603067, HQ603068, HQ603069, HQ603070, HQ603071, HQ603072, HQ603073, HQ603074, HQ603075, HQ603076, HQ603077, HQ603078, HQ603079, HQ603080, HQ603081, HQ603082.

2.3.2 Dating of ancient samples

Dating of samples was performed by collaborators. Sample ages were determined by direct 14C dating. These ages were calibrated using OxCal [203] (version 4.3) using the IntCal13 curve [204]. Table 2.1 shows the 14C age and standard deviation for each sample. This is followed by the median probability calibrated age before present (cal. bp). RISE386 was 14C dated twice, with ages (standard deviation) of 3,740 (33) and 3,775 (34); a rounded mean of 3,758 (34) was used for its calibration. DA29 was dated at 822 years using 14C and also at about 700 years using multi-proxy methods: the former date was used for consistency. The dates for DA119, DA222, RISE548, RISE556, RISE568 and RISE597 are best estimates, based on sample context.

2.3.3 Data and data processing

We analysed 101 Bronze Age samples published in Allentoft *et al.,* (2015) [41], 137 predominantly Iron Age samples published in de Barros Damgaard *et al.,* (2018) [44] and 66 additional samples from the Bronze Age. A total of 114.58×10^9 Illumina HiSeq 2500 sequencing reads were processed.

AdapterRemoval [205] (version 2.1.7) was used with its default settings to remove adaptors from all sequences, to trim N bases from the ends of reads and to trim bases with quality equal and smaller than 2. Reads were aligned against a human genome (GRCh38, https://www.ncbi.nlm.nih.gov/grc/human) using BWA [206] (version 0.7.15-r1140, mem algorithm). Reads that did not match the human genome were then mapped against the NCBI viral protein reference database containing 274,038

viral protein sequences (downloaded on 31 August 2016) using DIAMOND [207] (version 0.8.25). Protein matches were grouped into their corresponding viruses. Reads matching HBV were found in 25 samples.

The non-human reads from the samples that had more than three proteins matching HBV using DIAMOND were selected for a subsequent BLAST [25] (version 2.4.0) analysis. A BLAST database was made from Dataset 3, and samples were matched using BLASTn (with arguments `-task blastn -evalue 0.01`). Matching reads with bit scores greater than 50 for all samples (except DA222 (70) and DA45 (55)) were selected for subsequent processing. The number of reads selected from the BLAST matches, per sample, is shown in Table 2.1.

2.3.4 PCR confirmation and virus capture

PCR confirmation and virus capture was performed by Lasse Vinner. Capture probes were designed by Lasse Vinner and Anders J. Hansen. For a full description, please refer to [79].

Real-time PCR was established using primers and TaqMan probes as previously described [208], which were used to amplify a 91-base-pair amplicon of the HBV genome. Four samples were tested in total, two low-coverage (DA85, DA89) and two high-coverage (DA195, DA222) samples.

Fourteen samples with sufficient sample material were selected for virus capture (DA27, DA29, DA45, DA51, DA85, DA89, DA119, DA195, DA222, RISE254, RISE386, RISE416, RISE568 and RISE556).

The resulting reads were compared to Dataset 2 using BLASTn (with arguments `-task blastn -evalue 0.01`). Matching reads with bit scores greater than 50 for all samples (except DA222 (70) and DA45 (55)) were selected for subsequent processing. In total, 6,757 reads matched HBV in the capture data.

2.3.5 Sequence authenticity

The following evidence leads us to believe that the ancient HBV sequences are authentic and that the possibility of contamination can be excluded:

1. Standard precautions for working with aDNA were applied [63].
2. Sequences were checked for typical aDNA damage patterns using mapDamage [209] (version 2.0.6). Whenever sufficient amounts of data were avail-

able (>200 HBV reads), we found C → T mutations at the 5' end, typical of aDNA [67] (see Fig. 2.2a, c).

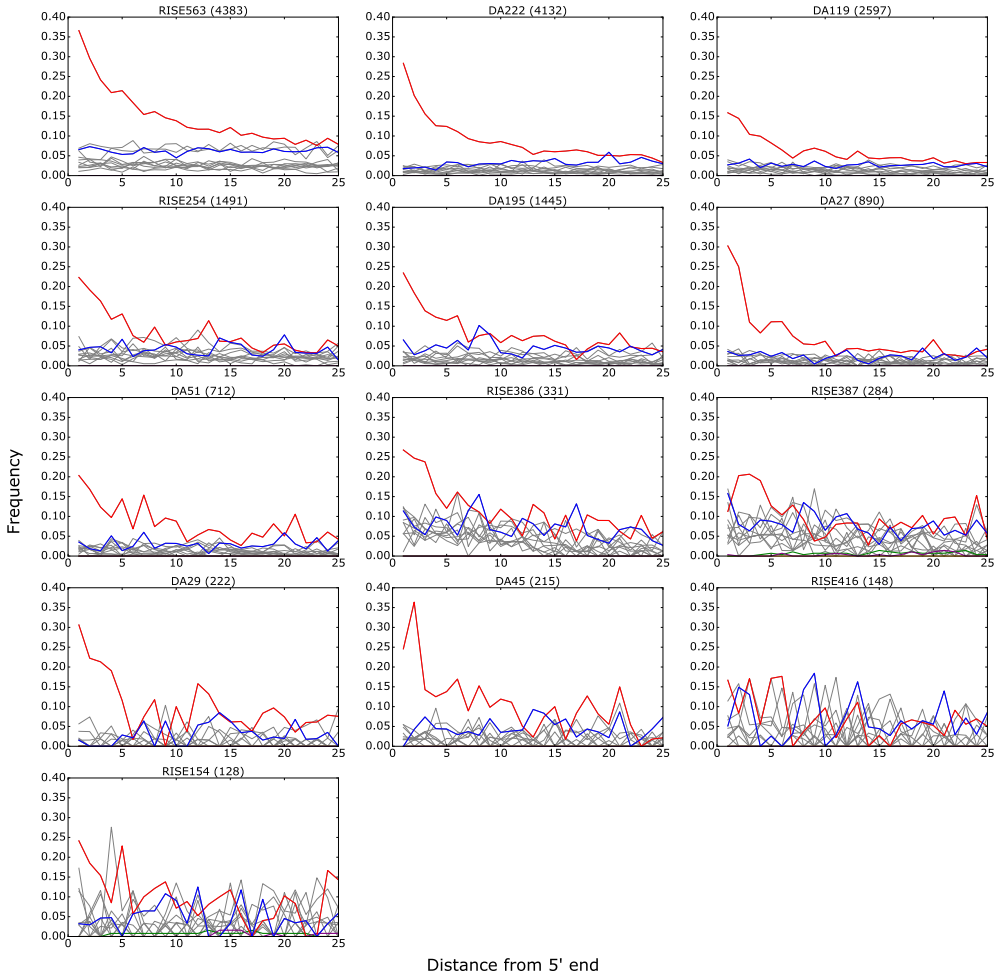
3. Capture was performed on DNA extracts of sample DA222, with and without pretreatment by uracil-specific excision reagent (USER) [210]. After USER treatment (3 h at 37 °C) of the aDNA extract, the damage pattern is eliminated (Fig. 2.2b).

4. As the ancient viruses are from three different HBV genotypes (A, B and D) and a clade in sister relationship to chimpanzee and gorilla HBVs, any argument that samples were contaminated would have to account for this diversity as well as the sequence novelty.

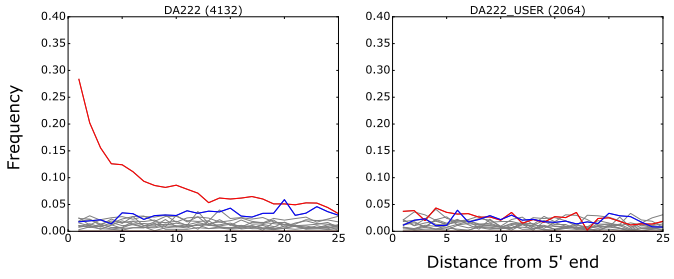
5. HBV sequences were identified in 25 of 304 analysed samples (Table 2.1), showing that the findings cannot be due to a ubiquitous laboratory contaminant.

6. Despite the low frequency of positive samples, we sequenced extraction blanks to provide additional evidence against the possibility that the HBV sequences stemmed from sporadic incorporation, amplification and sequencing of background reagent contaminants into the aDNA libraries. The negative extraction controls were amplified for 40 PCR cycles, and BLASTn was used to match the read sequences against Dataset 3, with the same parameters used for the ancient samples. Because the ancient HBV-positive reads used to assemble genomes all had bit scores of at least 50 (see section 2.3.3), we filtered the negative extraction control BLASTn output for reads with a bit score greater or equal to 45. No reads (out of 23 million) matched any HBV genome at that level.

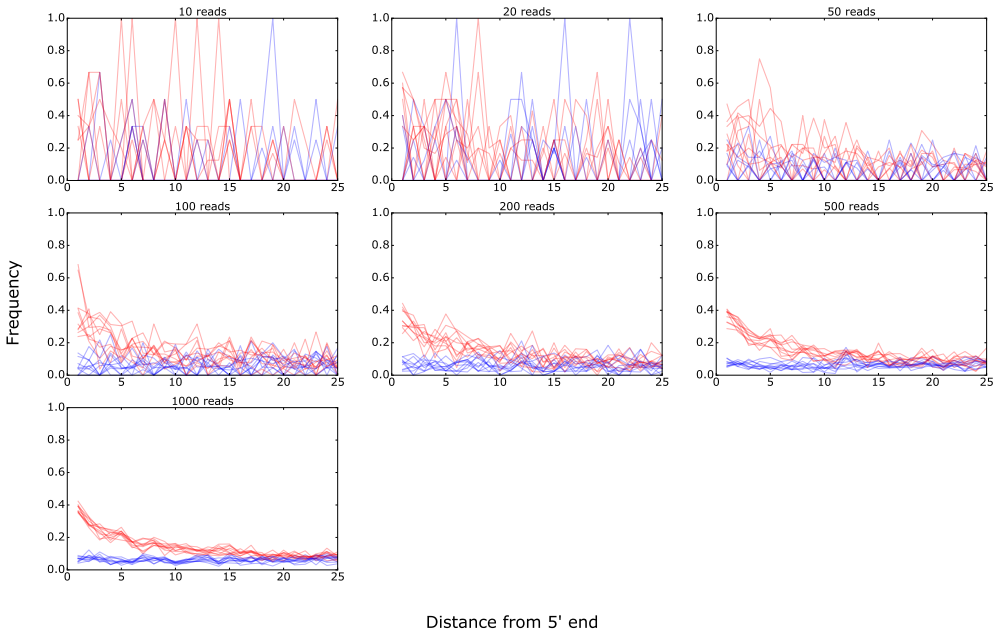
7. HBV is a blood-borne virus that is mainly transmitted by exposure to infectious blood and that does not occur in the environment [186], making contamination during archaeological excavation extremely unlikely.


RISE563 (4383)
DA222 (4132)
DA119 (2597)
RISE254 (1491)
DA195 (1445)
DA27 (890)
DA51 (712)
RISE386 (331)
RISE387 (284)
DA29 (222)
DA45 (215)
RISE416 (148)
RISE154 (128)
Frequency
Distance from 5' end

(a)


DA222 (4132)
DA222_USER (2064)
Frequency
Distance from 5' end

(b)



(c)

Figure 2.2: Ancient DNA damage patterns. The frequencies of the mismatches observed between the HBV reference sequences (Table 2.1) and the reads are shown as a function of distance from the 5' end. C $\rightarrow$ T (5') and G $\rightarrow$ A (3') mutations are shown in red and blue, respectively. All other possible mismatches are shown in grey. Insertions are shown in purple, deletions in green and clippings in orange. The count of reads matching HBV for each sample is shown in parentheses. **a) Damage patterns for RISE563, DA222, DA119, RISE254, DA195, DA27, DA51, RISE386, RISE387, DA29, DA45, RISE416 and RISE154. b) Damage patterns for DA222 without (left) and with (right) USER treatment. c) Damage patterns with 10, 20, 50, 100, 200, 500 and 1,000 reads sampled from RISE563, in which each opaque line corresponds to one replicate set of reads.** Only C $\rightarrow$ T (in red) and G $\rightarrow$ A (in blue) mutations are shown.

2.3.6 Consensus sequences

Reads from the original sequencing and from the capture were aligned to a reference genome (Table 2.1) in Geneious [211] (version 9) using medium sensitivity / fast and iterate up to 5 times. Because aDNA damage often clusters towards read termini [67], the resulting alignments were carefully curated by hand to remove non-matching termini of reads if the majority of the read showed a very good match with the reference sequence.

2.3.7 Genotyping

Performed by Terry Jones. All reads used to construct the ancient HBV consensus sequences were matched against the full NCBI nucleotide database (downloaded 28 December 2016) using BLAST. Of these reads, 97.5% had HBV as their top match. All ancient consensus sequences were matched against the full HBV genomes of dataset 4 with the Needleman-Wunsch algorithm [212], as implemented in EMBOSS50 (version 6.6.0.0). For each ancient sequence, the percentage of sequence identity with the most similar representative of each modern genotype and four NHP species is listed in Table 2.11. The Needleman-Wunsch algorithm was also used to calculate the pairwise sequence similarity between all ancient sequences (Table 2.12).

2.3.8 Recombination analysis

In order to gain an overview of possible recombination events present in dataset 2 and the 12 ancient sequences, a TreeOrder scan was performed as implemented in SSE (version 1.3) [213]. TreeOrder scan visualises changes of sequence positions in phylogenetic trees inferred from successive sections, or windows, of an alignment. A window of 500 base pairs and step size of 50 base pairs was used. Changes in tree order are shown if they have above 70% bootstrap support [235]. 100 bootstrap replicates were performed for each window. The woolly monkey HBV sequence (GenBank Accession Number: AF046996) was used as the outgroup. Subsequently, a GoupingScan was run as implemented in SSE (version 1.3) [213], using the sequences in dataset 2, after removing the sequence AB486012 (J). The remaining sequences in dataset 2 were assigned to 10 groups (genotypes A, B, C, D, E, F, G, H, and the Gibbon/Orang Utan clade, and the Chimpanzee/Gorilla clade). The GroupingScan measures how deeply embedded a test sequence (in this case, each of the 12 ancient sequences in turn) is in a group of reference sequences, measured

by the grouping score. A grouping score >0.5 is indicative that the test sequence falls within a particular group. The grouping score is calculated over successive windows of an alignment. We used a window size of 250 base pairs and a step size of 50 base pairs. 100 bootstrap replicates were performed. *The recombination analysis described below was performed by Terry Jones.* Furthermore, the recombination detection program (RDP4) [214] (version 4) was used to search for evidence of recombination within the 12 ancient sequences and a selection of 15 modern human and NHP sequences (Supplementary Methods in [79]). Recombination with HBV-RISE387 as the recombinant and HBV-DA51 as one parent, was suggested at positions 1567–2256, by seven recombination methods (RDP [214], GENECONV [215], BootScan [216], MaxChi [217], Chimaera [218], SiScan [219], and 3Seq [220]) with P values from 1.179×10^{-6} to 5.336×10^{-11} (Table 2.4, 2.5). The same recombination was suggested for all 4 ancient genotype A and two modern genotype A sequences. Graphical evidence of the recombination and the predicted break point distribution for sequences HBV-RISE386 and HBV-RISE387 from three methods (MaxChi, Bootscan and RDP) is shown in Fig. 2.7.

2.3.9 Initial maximum likelihood phylogeny and phylogenetic network

An initial maximum likelihood tree was generated to ascertain whether the ancient sequences fall within the primate HBV clades. Dataset 1 and the ancient sequences were aligned in MAFFT [221] (version 7). The maximum likelihood tree was constructed using PhyML [222] (version 20160116), optimizing topology, branch lengths and rates. We used a general time reversible (GTR) substitution model, with base frequencies determined by maximum likelihood, and a maximum likelihood-estimated proportion of invariant sites and 100 bootstraps (Fig. 2.3). Furthermore, a NeighbourNet phylogenetic network was constructed based on a MAFFT alignment of dataset 2 and the ancient sequences, using SplitsTree [223] with a GTR substitution model.

2.3.10 Dated coalescent phylogenies

To check for a temporal signal in the data, root-to-tip regressions and date randomization tests were performed. For the root-to-tip regression, input trees were calculated using dataset 2 with the addition of a woolly monkey sequence (GenBank Accession Number: AF046996) as an outgroup. Three phylogenetic algorithms were used;

neighbour joining, maximum likelihood (PhyML), and Bayesian (MrBayes [224] (version 3.2.5)) methods (Supplementary Figs. 1–3 in [79]). For maximum likelihood and Bayesian methods, root-to-tip distances (in substitutions per site) were extracted from optimized tree topologies (maximum likelihood and maximum clade credibility trees, respectively). For the neighbour joining method, root-to-tip distances were averaged over 1,000 bootstrap replicates. Root-to-tip distances were extracted using TempEst [225] (version 1.5). Regression analyses were performed with SciPy (version 0.16.0) [226]. For the date randomization tests, we used three different approaches to randomize tip dates. First, tip dates were randomized between all sequences in the phylogeny. Second, tip dates were randomized only among the ancient sequences presented here, as well as the Korean mummy sequence (GenBank accession number JN315779), while the modern sequences retained their correct ages. Third, dates were randomized within a clade. For each of the three approaches, we performed three independent randomizations. This resulted in a total of nine analyses, which were run for 100,000,000 generations each, under the relaxed log-normal clock model and coalescent exponential population prior. We also ran the same analyses under a strict clock and coalescent Bayesian skyline population prior, which were run for 20,000,000 generations. We used a GTR substitution model with unequal base frequencies, four gamma rate categories, estimated gamma distribution of rate variation and estimated proportion of invariant sites, as found by bModelTest [227] (version 1.0.4). None of the analyses using the relaxed clock converged (effective sample size <200). This is most probably because the misspecification of the dates leads to incongruence between the sequence and time information. Under the strict clock model, all runs converged, and none of the 95% HPD intervals of the root age overlapped between the randomized and the non-randomized runs, fulfilling the criteria for evidence of a temporal signal [228].

Dated phylogenies were estimated using BEAST2 [150] (version 2.4.4, prerelease). We used a MAFFT alignment of dataset 2. To account for the influence of recombination events on the dating of phylogenetic trees, we also performed all dating analysis on positions 1 – 1400 of a MAFFT alignment of dataset 2 after the exclusion of the genotype G sequences (hereafter referred to as 'partial alignment'), since the TreeOrder scan results (Fig. 2.4) indicate that that section of the alignment does not contain recombination events. Using bModelTest [227], we selected a GTR substitution model with unequal base frequencies, four gamma rate categories, estimated gamma distribution of rate variation and estimated proportion of invariant sites. Proper priors were used throughout. Path sampling, as implemented in BEAST2, was performed to select between a strict or relaxed log-normal clock and a coalescent constant, exponential or coalescent Bayesian skyline population prior

(Table 2.6). Log marginal likelihood values were compared using a Bayes factor test. A Bayes factor in the range of 3–20 implies positive support, 20–150 strong support and >150 overwhelming support [229]. The relaxed log-normal clock model in combination with a coalescent exponential population prior was favoured (Table 2.6). Note that there is a slight preference of the coalescent Bayesian skyline population prior instead of the coalescent exponential population prior under a relaxed log-normal clock (Bayes factor: 2.88) for the partial alignment. However, since this does not imply positive support [229], the same coalescent exponential population prior was used for the complete and the partial alignment (Table 2.6). Trees were inferred using the complete and partial alignments under all six combinations of clock models and population priors. The Markov chain Monte Carlo analysis was run until parameters reached an effective sample size >200, sampling every 2,000 generations. Convergence and mixing were assessed using Tracer [230] (version 1.6). For the final tree shown in figure 3.6, the final tree files were subsampled to contain 10,000 or 10,710 (for the relaxed log-normal clock, coalescent exponential population prior) trees, with the first 25% of samples discarded as burn-in. Maximum clade credibility trees were made using TreeAnnotator (version 2.4.4 prerelease).

To formally test the 'Out of Africa' hypothesis of HBV evolution, calibration points were tested using path sampling as implemented in BEAST2 for the full and partial alignment of dataset 2. Calibration points were constrained as follows: for the split of genotypes F and H, the most recent common ancestor (MRCA) of all genotype F and H sequences was constrained using a uniform (13,400:25,000) distribution, as this is the range of estimates for when the Americas were first colonized [231, 232]. For the split of subgenotype A3 in Haiti, the MRCA of FJ692598 and FJ692611 was constrained using a uniform (200:500) distribution, owing to the timing of the slave trade to Haiti [233]. For the split of C3 in Polynesia, the MRCA of X75656 and X75665 was constrained using a uniform (5,100:12,000) distribution, owing to the range of estimates for the MRCA of Polynesian populations [36, 234]. Calibration points were tested under both a relaxed log-normal clock, coalescent exponential population prior, and a strict clock, Bayesian skyline population prior.

2.4 RESULTS AND DISCUSSION

Shotgun sequence data were previously generated from 167 Bronze Age [41] and 137 predominantly Iron Age [44] individuals from central to western Eurasia with a sample age range of approximately 7.1–0.2 kya. We identified reads that matched the HBV genome in 25 samples (Table 2.1), spanning a period of almost 4,000 years, from several different cultures and with a broad geographical range (Fig. 2.1b, Table 2.1). Using TaqMan PCR, we tested two samples (DA195 and DA222) with high genome coverage and two samples (DA85 and DA89) with low genome coverage for the presence of HBV. The high-coverage samples tested positive, whereas the low-coverage samples tested negative (Table 2.3). This is consistent with shotgun sequencing being more effective than targeted PCR for analysing highly degraded DNA [58]. On the basis of the availability of sample material, libraries from 14 samples were selected for targeted enrichment (capture) of HBV DNA fragments (Supplementary Tables 1, 2 in [79]). This resulted in increased genome coverage and an average of a 2.4-fold increase in the number of HBV-positive reads (Table 2.2). We obtained 17.9–100% HBV genome coverage from the sequence data, with genomic depth ranging from 0.4x to 89.2x (Table 2.2). We selected 12 samples for phylogenetic analyses. Criteria for inclusion were at least 50% genome coverage and clear aDNA damage patterns after capture (Fig. 2.2a). For an initial phylogenetic grouping, we estimated a maximum likelihood tree using the ancient HBV genomes together with modern human, NHP, rodent and bat HBV genomes (dataset 1, see Methods). All ancient viruses fell within the diversity of Old World primate HBV genotypes, which includes all human and other great ape genotypes with the exception of human genotypes F and H (Fig. 2.3).

Sample	Location	Culture	^{14}C age (s.d.)	Median cal BP age/estimate	Approx. Age	Sex	Individual age	Sample type
DA27	Halvay 3, KAZ	Hun-Sarmatian	1641 (33)	1543	1610	M	N/D	N/D
DA29	Karasyur, KAZ	Golden Horde	849 (25)	755	822	M	N/D	N/D
DA45	Omnogobi, MNG	Xiongnu	2083 (27)	2053	2120	M	N/D	N/D
DA51	Keden, KYR	Saka	2220 (37)	2230	2297	M	N/D	N/D
DA85	Japyrik, KYR	Hun	1781 (46)	1702	1769	M	N/D	N/D
DA89	Berygavoya, KAZ	Turk	1315 (45)	1247	1314	M	N/D	N/D
DA119	Poprad, SVK	North Carpathian	N/D	1500	1567	M	N/D	N/D
DA195	Sandorfalva-Eperjes, HUN	Hungarian Scythian	2479 (35)	2578	2645	F	N/D	N/D
DA222	Butakty, KAZ	Karluk	N/D	1200–1000	1167	M	N/D	N/D
RISE99	L Bedinge, SWE	Nordic Late Neolithic	3713 (30)	4045	4112	M	Adult	N/D
RISE154	Szczepankowice, POL	Unetice	3522 (24)	3784	3851	F	Adult	tooth
RISE254	Szàzhalombatta-Földvàr, HUN	Vatya	3631 (29)	3942	4009	M	N/D	tooth
RISE391	Tanabergen II, KAZ	Sintashta	3612 (34)	3921	3988	F	30–35	tooth
RISE386	Bulanovo, RUS	Sintashta	3758 (34)	4121	4188	M	30–40	N/D
RISE387	Bulanovo, RUS	Sintashta	3822 (33)	4215	4282	N/D	N/D	tooth
RISE416	Nerquin Getashen, ARM	Middle Bronze Age	3259 (40)	3488	3555	M	Infant	N/D
RISE478	Schöngeising, GER	Bronze Age	3164 (34)	3391	3458	M	Adult	N/D
RISE509	Bateni, RUS	Afanasievo	4186 (27)	4728	4795	F	20–25	tooth
RISE533	Arano, ITA	Early Bronze Age	3657 (37)	3980	4047	N/D	N/D	tooth
RISE548	Temrta IV, RUS	Yamnaya	N/D	5000–4300	4717	M	Senilis (55+)	tooth
RISE554	Afontova Gora, RUS	Late Bronze Age hunter-gatherer	2782 (30)	2879	2946	M	N/D	tooth
RISE556	Manching-Oberstimm, GER	Bell Beaker	N/D	4500–4000	4317	F	Adult	tooth
RISE563	Osterhofen-Altenmarkt, GER	Bell Beaker	3955 (35)	4421	4488	M	Adult	tooth
RISE568	Brandỳsek,CZE	Early Slav	N/D	1350–1150	1317	F	Infant	tooth
RISE597	Turlojiske, LTU	Late Bronze Age	N/D	3500–3000	3317	M	Adult	tooth

Table 2.1: Overview of samples with reads matching HBV. Cal BP: calibrated years before present (taken as 1950 AD); ND: not determined (samples for which dating was not performed, or for which osteological sex was undetermined).

Sample	Proteins Diamond	Reads Diamond	Reads BLASTn	Bit score cut-off	Captured reads BLASTn	Reads aligned in Geneious	Reference	Coverage of consensus before capture	Coverage depth before capture	Coverage of consensus	Coverage depth
DA27	7	309	890	50	8	890	DQ315779	86.11%	15.3x	90.00%	14.3x
DA29	7	149	179	50	44	222	DQ315779	84.22%	3.7x	87.50%	4.8x
DA45	4	29	110	55	118	215	AB073858	67.03%	1.9x	87.20%	4.3x
DA51	7	54	127	50	595	712	DQ315779	69.42%	1.9x	99.20%	14.5x
DA85	4	14	27	50	6	N/D	DQ315779	29.80%	0.7x	35.80%	0.5x
DA89	7	5	51	50	0	N/D	DQ315779	17.90%	1.0x	17.90%	1.0x
DA119	7	155	385	50	2272	2597	AY738142	86.00%	7.2x	98.80%	53.1x
DA195	7	353	694	50	774	1445	AY738142	99.41%	12.6x	99.90%	29.2x
DA222	7	2269	2855	70	1318	4132	DQ315779	98.81%	61.2x	100%	89.2x
RISE99	2	2	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D
RISE154	7	17	137	50	N/D	128	AB073858	N/D	N/D	57.20%	2.0x
RISE254	7	158	211	50	1304	1491	DQ315779	52.58%	5.2x	99.00%	36.6x
RISE391	7	18	42	50	N/D	N/D	AB073858	N/D	N/D	34.60%	0.9x
RISE386	7	94	240	50	140	331	AY738142	86.28%	4.2x	97.80%	7.0x
RISE387	7	154	344	50	N/D	284	AY738142	N/D	N/D	86.60%	6.2x
RISE416	6	4	14	50	137	N/D	AY738142	10.60%	0.3x	75.80%	3.5x
RISE478	3	2	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D
RISE509	2	2	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D
RISE533	6	3	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D
RISE548	5	5	19	50	N/D	N/D	AB073858	N/D	N/D	17.90%	0.5x
RISE554	2	1	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D
RISE556	2	2	N/D	50	32	N/D	AY738142	N/D	N/D	28.00%	0.6x
RISE563	7	1499	4817	50	N/D	N/D	AB188243	N/D	N/D	100%	79.3x
RISE568	7	6	7	50	9	N/D	AB073858	N/D	N/D	33.90%	0.4x
RISE597	3	1	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D	N/D

Table 2.2: Mapping statistics of samples with reads matching HBV. 'Proteins Diamond' denotes the number of HBV proteins matched by HBV sequencing reads using DIAMOND, 'Reads Diamond', the number of sequencing reads that matched HBV using DIAMOND; 'Reads BLASTn', the number of sequencing reads that matched HBV; 'Captured reads BLASTn', the number of reads from the capture that matched HBV using BLASTn; 'Bit score cut-off', the bit score cut-off above which matching reads were used to form consensus sequences; 'Reads aligned in Geneious', the number of reads matching against reference sequence in Geneious, 'Reference' gives the accession number of the sequence that was used to map the reads against in Geneious to make the consensus, 'Coverage before capture' and 'Coverage depth before capture' show the coverage and depth of the consensus sequence, excluding the reads from the capture. 'Coverage consensus' and 'Coverage depth', the percentage of the consensus genome covered by matching reads and average depth of coverage across the reference genome as reported by Geneious. When reading sample information across a row, an N/D (not determined) cell will be encountered when processing on that sample was concluded, either owing to too few matching reads or consensus coverage less than 50%. Samples for which no capture was done (RISE563, RISE387, RISE154, RISE391, RISE548, RISE533, RISE99, RISE478, RISE509, RISE554, RISE597) and samples for which no pre-capture estimates are available due to the low number of matching reads (RISE556) are also denoted as 'N/D'.

Sample	% of HBV reads per sample	CT-value	HBV DNA qPCR conclusion	Remarks
DA222	7.45E-05	31	Positive	M
DA195	2.82E-05	40.52	Positive	F
DA85	6.41E-07	>45	Negative	
DA89	2.85E-06	>45	Negative	
DA351 (neg.)	NA	>45	Negative	
Negative ancient samples (n=4)	NA	>45	Negative	
NTC (n=8)	NA	>45	Negative	
Enriched HBVpos library pool 1	NA	>40	Positive	
Enriched HBVpos library pool 2	NA	35.39	Positive	
Enriched HBVneg library pool 3	NA	>45	Negative	
Enriched HBVneg library pool 4	NA	>45	Negative	

Table 2.3: TaqMan PCR results. Four extracts from samples with HBV reads were selected for TaqMan PCR confirmation: two with a large proportion of HBV reads (DA222 and DA195), two with a small proportion of HBV reads (DA85 and DA89) and one with no HBV reads (DA351). HBV was detected in extracts from DA222 and DA195, whereas the three low- and zero-read samples were negative, as were all no-template controls. *Table contributed by Lasse Vinner.*

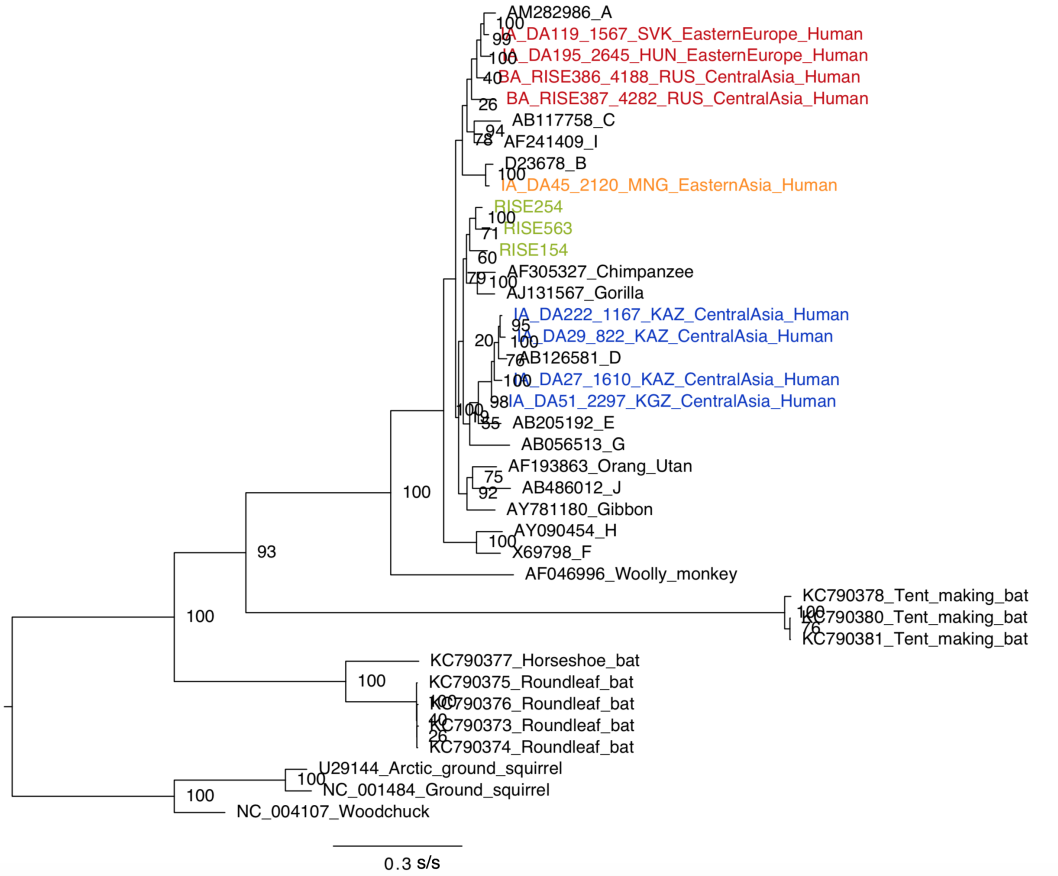


Figure 2.3: *Hepadnaviridae* maximum likelihood tree. This figure shows 26 *Orthohepadnaviridae* sequences (Dataset 1, see Methods), including the ancient HBV sequences. Ancient genotype A sequences are shown in red, the ancient genotype B sequence in orange, ancient genotype D sequences in blue and novel genotype sequences in green. The tree was constructed in PhyML [222], optimizing for topology, branch lengths and rates, with 100 bootstraps (see Methods). The x-axis denotes substitutions per site (s/s).

Recombination is known to occur in HBV [235]. For an overview of possible recombination events, we visualised changes in tree order for subsequent sections of an alignment of HBV sequences using TreeOrder scan [213]. Figure 2.4 shows that changes to the phylogenetic relationship between genotypes occur mainly between alignment positions ∼1400 – 3221, except for genotype G, which also changes tree order around position 300. The ancient sequences group with a single genotype across the whole genome, with the exception of HBV-RISE387 and HBV-DA29. To gain additional information about possible recombination events between the ancient and the modern sequences, we performed a GroupingScan [213] (Fig. 2.5). The GroupingScan computes a grouping score for successive windows along the alignment as a measure of how embedded a test sequence (here, each of the 12 ancient sequences in turn) is in a group of non-recombinant sequences (corresponding to the modern HBV genotypes A – H, and non-human primate sequences). Using a window size of 250 base pairs, the GroupingScan highlighted possible recombination events in HBV-RISE387 (with genotype G), -RISE386 (with genotype D), -DA195 (with genotype D), -DA29 (with genotype E), and -DA45 (with gibbon and orang utan sequences). Furthermore, sequences HBV-RISE154, -RISE254, and -RISE563 appear to have contributions from a variety of different genotypes (D, E, and G) as well as the non-human primate sequences. This may reflect the absence of reference sequences closely related to HBV-RISE154, -RISE254, and -RISE563 in the dataset. The possible recombination events in HBV-DA45 and HBV-DA195 are not present if a larger window size of 500 base pairs is used (Fig. 2.6). Since the GroupingScan measures how embedded a test sequence is in a group of non-recombinant sequences, it can only compare a single test sequence to a group of reference sequences. We also used seven different recombination detection programs implemented in RDP4 [214], in order to investigate whether a single ancient sequence has recombined with an ancient or modern sequence. We do not find evidence for a recombination event between HBV-DA45 and a gibbon or orang utan sequence or between HBV-DA29 and modern genotype E. Should the recombination event with genotype E indicated by the GroupingScan be correct, the age of HBV-DA29 (∼822 years old) together with the fact that a similar recombination event is not observed in the older ancient genotype D sequences, indicate that the recombination event may have taken place around ∼800 years ago. Using RDP4, we find evidence of a recombination event between HBV-DA51 and an unknown parent which formed the ancient genotype A sequences. Although a recombination event between HBV-DA51 and the ancient genotype A sequences cannot have occurred owing to sample ages, the logical interpretation is that an ancestor of HBV-DA51 was involved in the recombination. The same recombination is also suggested for the two modern genotype A sequences that were included

in the analysis. The ancient genotype B (HBV-DA45), a modern genotype B and two modern genotype C sequences were not similarly flagged, which suggests that the possible recombination occurred after genotypes A, B, and C had diverged. The predicted recombination break points (Table 2.4, 2.5 and Fig. 2.7) correspond closely to the polymerase gene. It is therefore possible that the polymerase from an unknown parent and the remainder of the genome from an HBV-DA51 ancestor recombined to form the now-ubiquitous genotype A about 7.4–9 kya (Fig. 3.6, Table 2.7 and Methods). Similar recombination events that involved the creation of genotypes E, G and a currently circulating B/C recombinant have previously been identified [235].

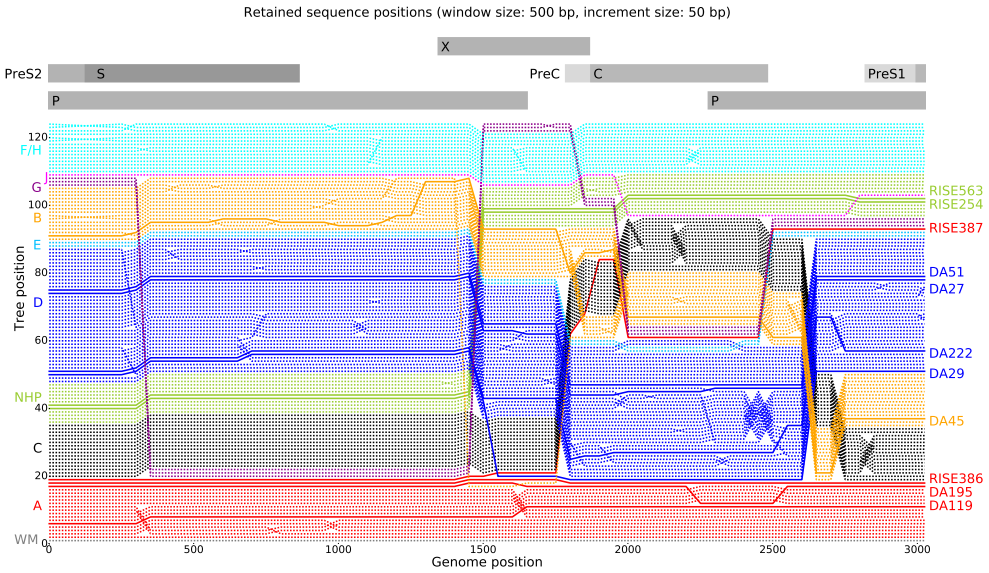


Figure 2.4: TreeOrder scan of modern and ancient HBV sequences. The x-axis shows the genome position, the y-axis the relative position of sequences in a phylogenetic tree, computed over successive windows of 500 base pairs, incrementing by 50 base pairs. Changes in tree order with >70% bootstrap support are shown. The woolly monkey sequence (NCBI accession number: AF046996) was used as the outgroup. Modern and ancient sequences are shown as dotted and continuous lines, respectively. Genotypes are labelled on the left in their respective colours, 'WM' indicates the woolly monkey sequence, and 'NHP' indicates sequences from non-human primates. Ancient sequences are labelled on the right. The grey bars at the top of the figure indicate the positions of genes in the HBV genome.

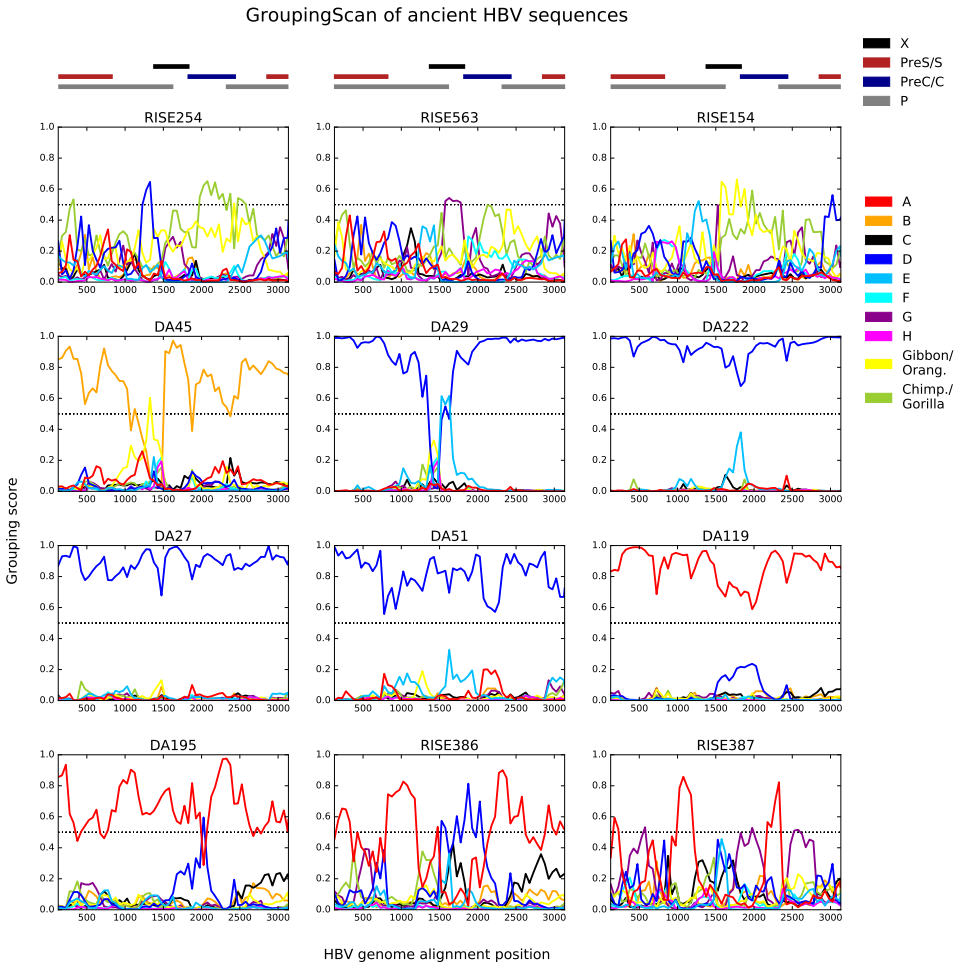


Figure 2.5: GroupingScan of ancient HBV sequences. The x-axis shows the genome position, the y-axis the grouping score, computed over successive windows of 250 base pairs, incrementing by 50 base pairs. The grouping score indicates how deeply embedded a test sequence is in a group of reference sequences. A grouping score >0.5 (indicated by a dotted line) indicates grouping of the test sequence within a specific group. The horizontal bars at the top of the figures on the first line indicate the positions of genes in the HBV genome.

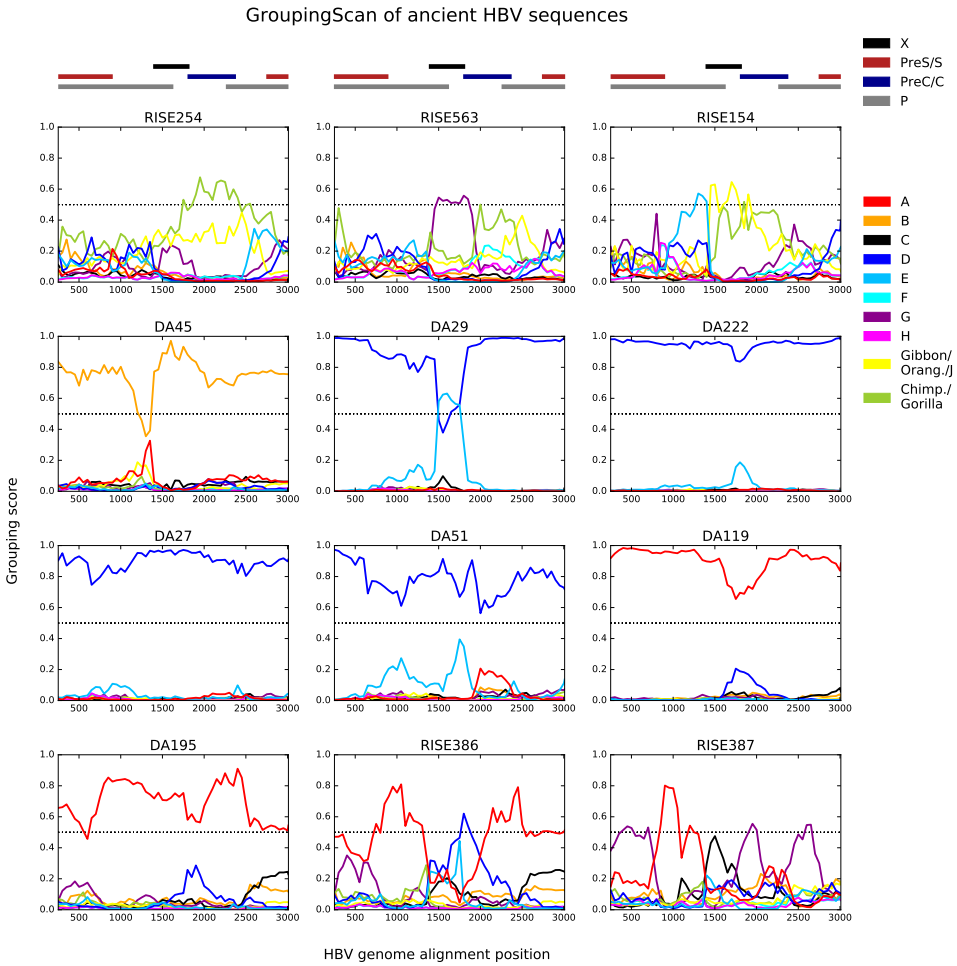


Figure 2.6: GroupingScan of ancient HBV sequences. The x-axis shows the genome position, the y-axis the grouping score, computed over successive windows of 500 base pairs, incrementing by 50 base pairs. The grouping score indicates how deeply embedded a test sequence is in a group of reference sequences. A grouping score >0.5 (indicated by a dotted line) indicates grouping of the test sequence within a specific group. The horizontal bars at the top of the figures on the first line indicate the positions of genes in the HBV genome.

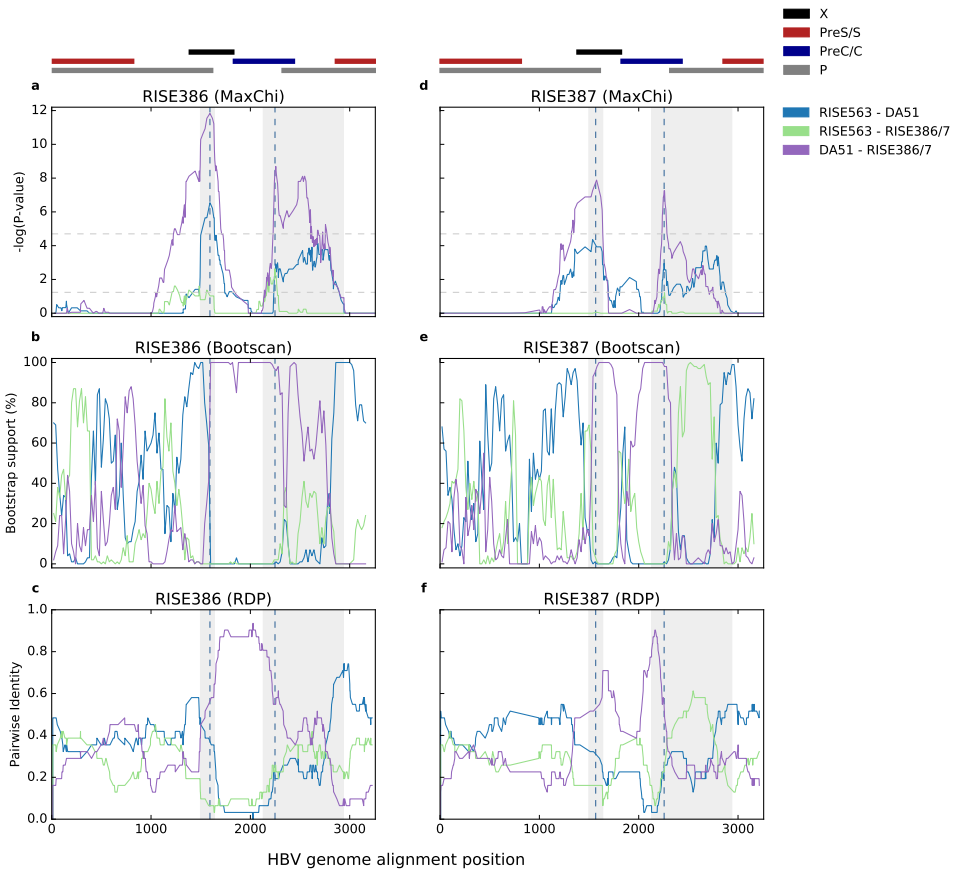


Figure 2.7: Genotype A recombination break-point evidence. RDP4 [214] was used to analyse the set of 12 ancient sequences plus a representative set of 15 modern human and NHP sequences. The seven recombination programs used by RDP4 suggested that all genotype A sequences are recombinants, with the genotype D sequence HBV-DA51 as the minor parent and an unknown major parent. The obvious interpretation is that recombination formed an ancestor of the oldest sequences, evidence of which is still present in the less-ancient and the modern representatives. The figure shows the graphical evidence and predicted recombination break-point distribution for the two oldest genotype A sequences, HBV-RISE386 and HBV-RISE387, according to three of the RDP4 methods (MaxChi, Bootscan and RDP). In all subplots, the predicted location of the break points is shown as a dashed vertical line and the surrounding grey area shows the 99% confidence interval for the break point. Subplots on the same row share their y-axis and those in the same column share their x-axis. The horizontal bars at the top of the figures on the first line indicate the positions of genes in the HBV genomes. **a) HBV-RISE386 analysed by MaxChi. b) HBV-RISE386 analysed by Bootscan. c) HBV-RISE386 analysed by RDP. d) HBV-RISE387 analysed by MaxChi. e) HBV-RISE387 analysed by Bootscan. f) HBV-RISE387 analysed by RDP.** *Figure made by Terry Jones.*

Method	Number of sequences	Average P value
RDP	6	1.641×10^{-6}
GENECONV	6	2.450×10^{-6}
BootScan	6	1.179×10^{-6}
MaxChi	6	3.494×10^{-6}
Chimaera	6	2.332×10^{-6}
SiScan	6	5.336×10^{-11}
3Seq	6	$7.319 \text{Ex} 10^{-8}$

Table 2.4: The P values assigned to the predicted genotype A recombination by the seven methods used by RDP4 [214], in the order given by RDP4. The number of sequences in which the recombination was predicted is always six, corresponding to the four ancient and two modern genotype A sequences. *Table made by Terry Jones*.

Sample	Begin	Begin 99% Confidence Interval	End	End 99% Confidence Interval
RISE387	1567	1498–1638	2256	2130–2937
RISE386	1593	1498–1638	2248	2130–2937
DA195	1633	1498–1638	2247	2130–2937
DA119	1633	1498–1638	2247	2130–2937
KJ854705	1642	1498–1638	2247	2130–2937
LC074724	1622	1498–1638	2256	2130–2937

Table 2.5: The predicted start and end break points for each of the six genotype A sequences. Sequences are ordered from oldest to youngest. The 99% confidence intervals for the start and end points are shown, and are identical for all sequences. The predicted break points are close to the boundaries of the polymerase. For example, for the modern genotype A sequence LC074724, the polymerase is found in regions 1–1623 and 2307–3221 and the predicted break points are 1622 and 2256. If recombination formed an HBV-RISE387/HBV-RISE386 ancestor, it is possible that the entire polymerase gene was contributed by one parent. *Table made by Terry Jones*.

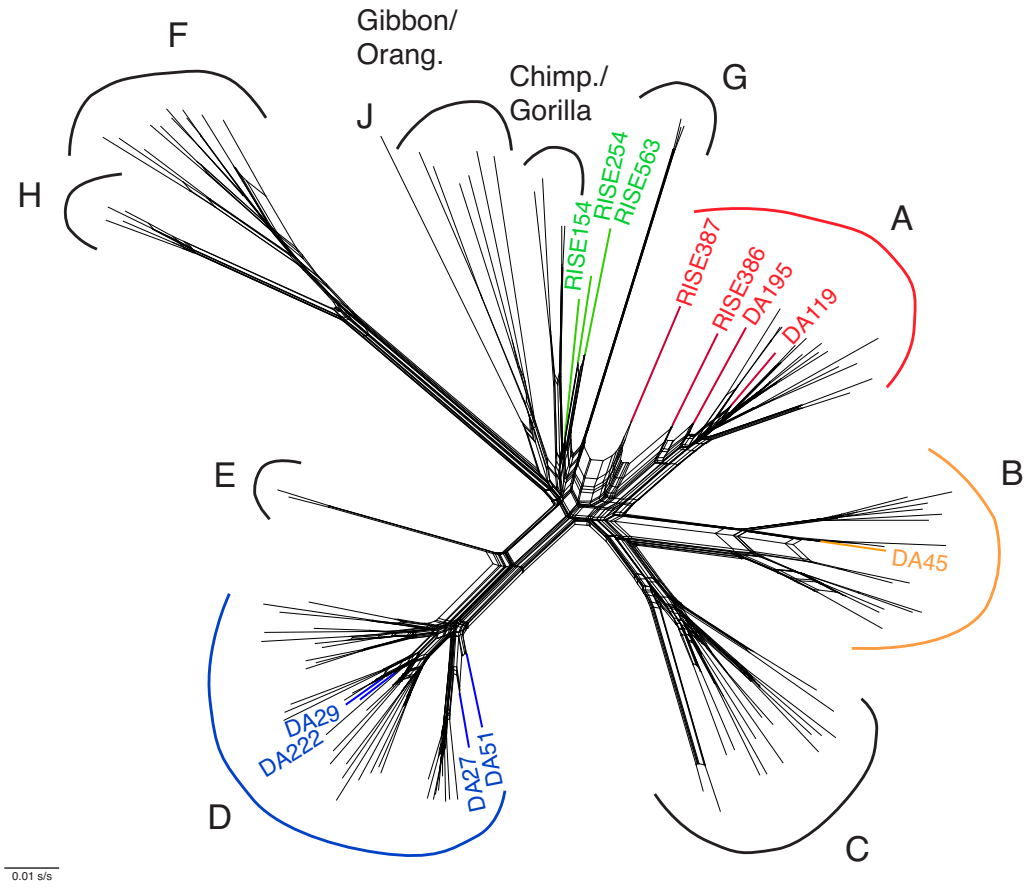


Figure 2.8: NeighborNet phylogenetic network. The phylogenetic network was constructed using SplitsTree [223], using a GTR substitution model. Ancient genotype A sequences are shown in red, ancient genotype B sequences in orange, ancient genotype D sequences in blue and novel genotype sequences closest to NHP HBV in green. The scale bar denotes substitutions per site (s/s).

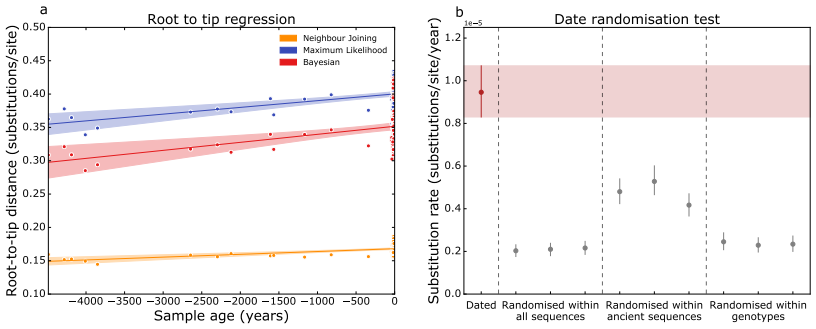


Figure 2.9: Root-to-tip regression and date randomization tests. a) Regression of root-to-tip distances and ages performed in SciPy. Branch lengths were extracted using TempEst [225] from trees inferred using neighbour joining, maximum likelihood and Bayesian methods. Shaded areas show 95% confidence intervals. Slopes are 1.01×10^{-5}, 1.20×10^{-5} and 4.21×10^{-6}, and correlation coefficients are 0.45 (R^2 = 0.2), 0.36 (R^2 = 0.13) and 0.51 (R^2 = 0.26), for maximum likelihood, Bayesian and neighbour joining trees, respectively. **b) Date randomization tests under the strict clock model.** The median and 95% HPD interval for the substitution rates are given. The rate for the correctly dated tree is shown in red. Dates were randomized within all sequences, within the ancient sequences only, and within each genotype. None of the 95% HPD intervals for the randomized runs overlaps with the 95% HPD intervals for the correctly dated runs, suggesting the presence of a temporal signal in the data.

For detailed phylogenetic analyses, we used a set of 112 reference human and NHP HBV sequences (dataset 2, see Methods). A NeighborNet phylogenetic network based on these reference sequences and the 12 ancient sequences was constructed, to visualise the relationship between sequences while taking into account any recombination events (Fig. 2.8). Four sequences (HBV-RISE387, -RISE386, -DA195, and -DA119) group within or basal to genotype A, four group with genotype D (HBV-DA51, -DA27, -DA222, and -DA29), one with genotype B (HBV-DA45), and three with the Chimpanzee and Gorilla sequences (HBV-RISE154, -RISE254, and -RISE563). Regression of root-to-tip genetic distances against sampling dates, as well as date randomization tests, showed a clear temporal signal in the data (Fig. 2.9 and Supplementary Figs. 1–3 in [79]), suggesting that molecular clock models can be applied. A dated coalescent phylogeny was constructed using BEAST2 [150] (Fig. 3.6). The molecular clock was calibrated using tip dates. Strict and relaxed log-normal molecular clocks were tested with coalescent constant, exponential and Bayesian skyline population priors (Table 2.6). Model comparisons favoured a relaxed molecular clock model with log-normally distributed rate variation and a coalescent exponential population prior (Table 2.6). The median root age of the resulting tree is estimated to be 11.6 kyr (95% highest posterior density (HPD) interval:

8.6–15.3 kyr) and the median clock rate is 1.18×10^{-5} substitutions per site per year (95% HPD interval: 9.21×10^{-6}–1.45×10^{-5} substitutions per site per year). Under a strict molecular clock, a coalescent Bayesian skyline population prior was favoured, in which case the median root age is 15.6 kyr (95% HPD interval: 13.7–17.8 kyr) and the median substitution rate is 9.48×10^{-6} substitutions per site per year (95% HPD interval: 8.3×10^{-6}–1.07×10^{-5} substitutions per site per year) (Tables 2.6, 2.7, 2.8).

Since recombination can affect the topology and dating of phylogenetic trees, we also inferred dated coalescent trees for the non-recombinant section of the genome from position 1 – 1400 and with the removal of genotype G (hereafter referred to as 'partial alignment'), as indicated by TreeOrder scan (Fig. 2.4). Most recent common ancestor (MRCA) age for the root as well as for different genotypes and sub-clades for the full and partial alignments are shown in Fig. 2.11a. We do not observe differences in MRCA age inferred from the complete and the partial alignments where the 95% highest posterior density intervals do not overlap, except for genotypes B and C (Tables 2.7, 2.8). Figures 2.11b–d show schematic trees for the complete alignment (Fig. 2.11b), and the the partial alignment (using a strict clock (Fig. 2.11c), and a relaxed clock (Fig. 2.11d)), indicating that changes in genotype B and C MRCA age are associated with differences in the tree topology. The substitution rates inferred under different clock models and population priors for the complete and partial alignments do not differ such that the 95% highest posterior density intervals no longer overlap (Table 2.8). Thus, apart from the MRCA age of genotypes B and C, recombination does not seem to affect the inferred dates and substitution rates.

Under the complete and partial alignment, and all model parameterizations used here, the substitution rate that we find (Table 2.8) is lower than rates estimated from phylogenies built using either modern heterochronous sequences [37] or sequences from mother-to-child transmissions [236] but higher than rates inferred using external calibrations based on human migrations [36]. A lower rate is consistent with previous work [237] in which it was shown that, although mutation rates may be high, mutations within an individual often revert back to the genotype consensus and thus rarely lead to long-term sequence change. It is also consistent with the time-dependent rate phenomenon, observed for many viruses, which suggests that short-term evolutionary rates are higher than long-term rates [162]. The ancient HBV genome data enable us to formally evaluate hypotheses concerning HBV origins using path sampling of calibrated phylogenies based on appropriate external divergence date assumptions. We tested several calibration points that would be implied by a co-expansion of HBV with humans after leaving Africa for support of congruence between migrations and

geographical locations of HBV clades [36]. We find weak evidence for the split of the F and H clade occurring between 13.4 and 25.0 kya under a strict, but not a relaxed clock model for the complete alignment and strong support under the strict and the relaxed clock for the partial alignment. We do not find support for the divergence of subgenotype C3 strains between 5.1 and 12.0 kya (hypothesized to have led to its distribution in different regions of Polynesia [36]). Using the complete alignment, we do not find support for divergence of Haitian A3 strains from other genotype A strains between 0.2 and 0.5 kya under either strict or relaxed clock models, but weak support for the partial genome under a strict clock (Table 2.9).

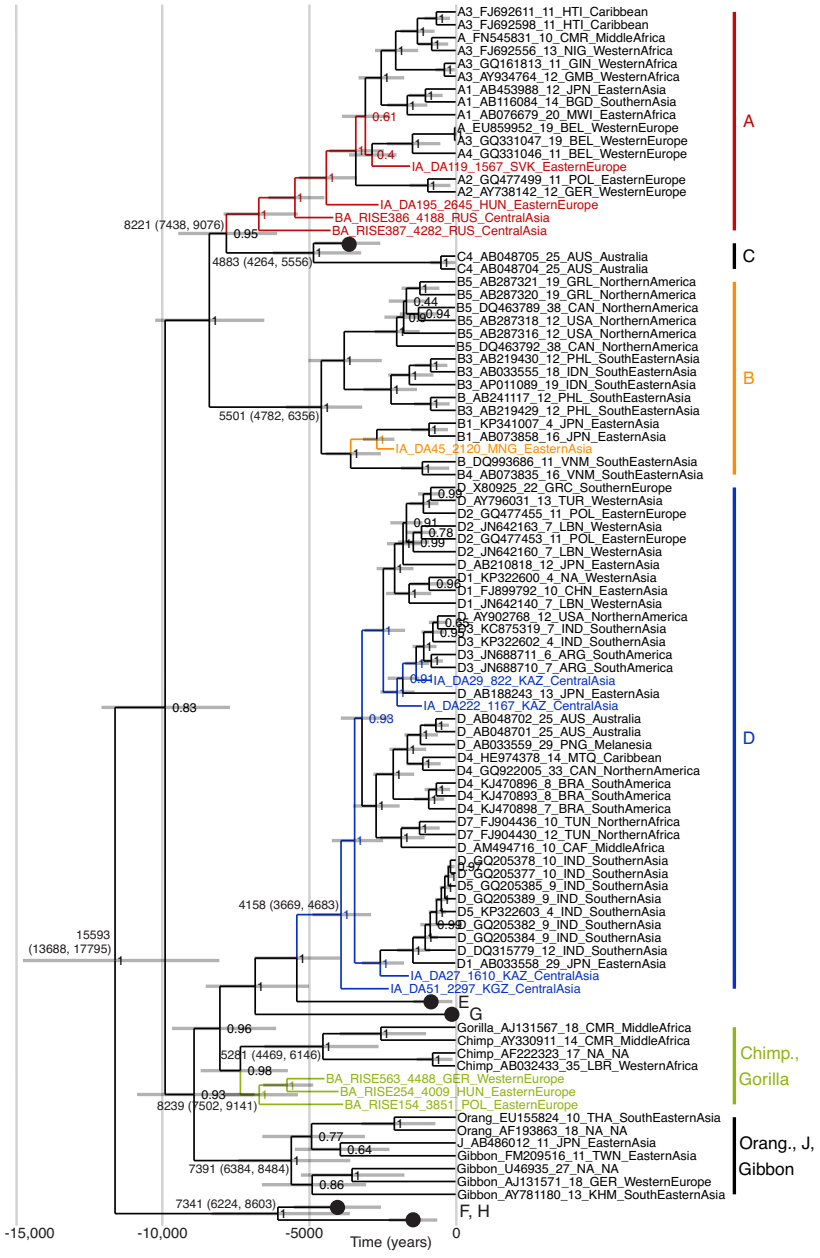


Figure 2.10: Dated maximum clade credibility tree of HBV. A log-normal relaxed clock and coalescent exponential population prior were used. Grey horizontal bars indicate the 95% HPD interval of the age of the node. Larger numbers on the nodes indicate the median age and 95% HPD interval of the age (in parentheses) under a strict clock and Bayesian skyline tree prior. Clades of genotypes C (except clade C4), E, F, G and H are collapsed and shown as dots. The figure includes a possible tenth genotype, J, based on a single human isolate. Taxon names for ancient samples indicate era (BA, Bronze Age; IA, Iron Age or later), sample name, sample age in years, ISO 3166 three-letter abbreviation of country of sequence origin, and region of sequence origin. Taxon names for modern samples indicate human genotype or subgenotype or host species if non-human, GenBank accession number, sample age in years, ISO 3166 three-letter abbreviation of country of sequence origin, and region of sequence origin. The x-axis shows time into the past, in years.

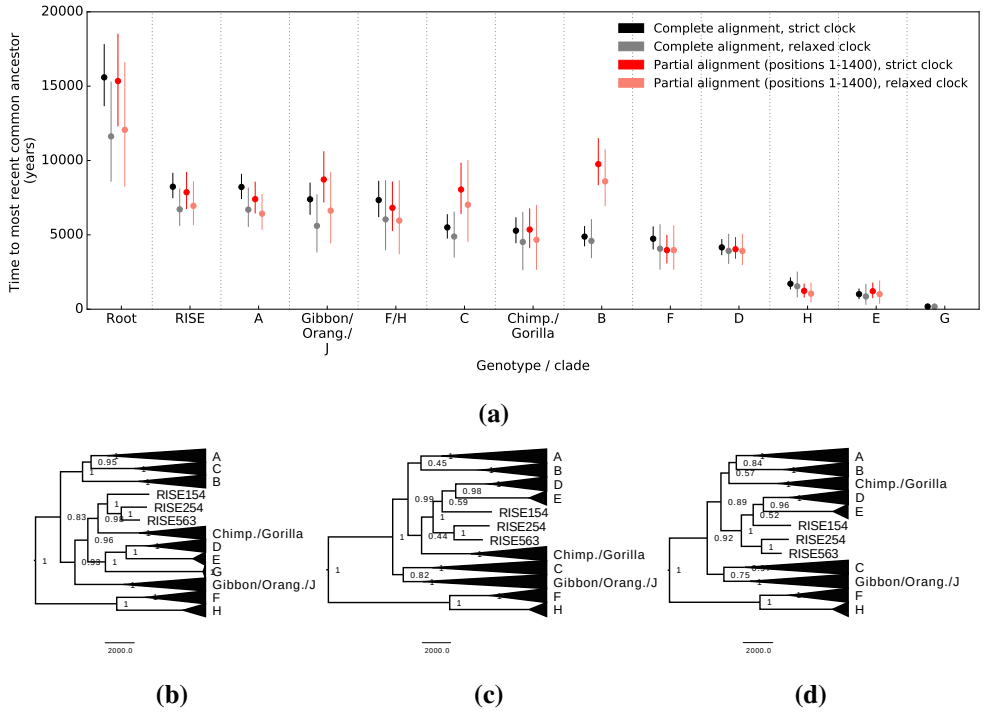


Figure 2.11: Time to most recent common ancestor and tree topology for different genotypes and parts of the HBV genome. a) Time to most recent common ancestor (shown on the y-axis) for different genotypes or clades (shown on the x-axis). **b – d)** Schematic trees showing the tree topologies inferred using the complete genome, a relaxed log-normal clock (b), the partial alignment from positions 1 – 1400, leaving out genotype G due to recombination, using a strict clock and coalescent Bayesian skyline population prior (c), or a relaxed log-normal clock and a coalescent exponential population prior (d). Scale bars in figures b – d indicate time in years.

In the dated coalescent phylogeny, four ancient sequences (from youngest to oldest: HBV-DA119, HBV-DA195, HBV-RISE386 and HBV-RISE387) group with genotype A. The first three are well within the 7.5% nucleotide divergence criterion that was used to delimit membership in HBV genotypes, and HBV-RISE387 is right on this limit (7.51%) [196] (Table 2.11). The three oldest samples lack a six-nucleotide insertion at the carboxyl end of the core gene (C) that is present in all modern genotype A viruses [193] (Table 2.10). HBV-RISE387 encodes a stop codon in its pre-core peptide that would have ablated the expression of the immune modulator HBe antigen (HBeAg), a phenomenon that is known to occur in modern HBV infections (Table 2.10). This characteristic viral mutant is usually found in chronic HBV carriers who seroconverted from HBeAg to anti-HBe. RISE386 and RISE387 have archaeologically been dated to only about 100 years apart and both come from the Bulanovo

Model (clock, population prior)		Strict, Constant	Strict, Exponential	Strict, Skyline	Relaxed-log., Constant	Relaxed-log., Exponential	Relaxed-log., Skyline
	Likelihood	-50556.68	-50503.88	-50499.63	-50500.28	-50324.79	-50334.43
Strict, Constant	-50556.68	0	52.8	57.048	56.4	231.88	222.24
Strict, Exponential	-50503.88		0	4.25	3.6	179.08	169.44
Strict, Skyline	-50499.63			0	-0.65	174.83	165.2
Relaxed-log., Constant	-50500.28				0	175.48	165.85
Relaxed-log., Exponential	-50324.79					0	-9.64
Relaxed-log., Skyline	-50334.43						0

(a) Path sampling results computed using the complete alignment of dataset 2.

Model (clock, population prior)		Strict, Constant	Strict, Exponential	Strict, Skyline	Relaxed-log., Constant	Relaxed-log., Exponential	Relaxed-log., Skyline
	Likelihood	-19461.4298	-19434.7391	-19423.04876	-19419.85918	-19385.12528	-19382.24344
Strict, Constant	-19461.43	0	26.69	38.38	41.57	76.30	79.19
Strict, Exponential	-19434.74		0	11.69	14.88	49.61	52.49
Strict, Skyline	-19423.05			0	3.19	37.92	40.81
Relaxed-log., Constant	-19419.86				0	34.73	37.62
Relaxed-log., Exponential	-19385.12					0	2.88
Relaxed-log., Skyline	-19382.24						0

(b) Path sampling results computed using the partial alignment (positions 1 – 1400) of dataset 2.

Table 2.6: Results of testing different clock models and population assumptions to be used for dated phylogenies. Different combinations of clock models and population assumptions are compared by comparing their log marginal likelihoods. Abbreviations are as follows: 'Strict': strict clock, 'Relaxed-log.': relaxed log-normal clock, 'Constant': coalescent constant population prior, 'Exponential': coalescent exponential population prior, 'Skyline': coalescent Bayesian skyline population prior, 'likelihood': log marginal likelihood estimate. Log marginal likelihoods (shown in the second row and column of Table a and b) were computed using path sampling, as implemented in BEAST2 [150]. Numbers in columns and rows 3 – 8 in Tables a and b are Bayes factors, computed by subtracting the log marginal likelihood of the row model from the log marginal likelihood of the column model. According to previous work [229], a Bayes factor in the range of 3–20 implies positive support, 20 – 150 strong support and >150 overwhelming support. Positive Bayes factors indicate support for the column model, and negative Bayes factors indicate support for the row model. **a) Path sampling results for the complete genome alignment of dataset 2. b) Path sampling results for the partial alignment (positions 1 – 1400 of dataset 2).**

Clade	Complete, MRCA age in years (95% HPD interval), Strict, Skyline	Partial, MRCA age in years (95% HPD interval), Strict, Skyline	Complete, MRCA age in years (95% HPD interval), Relaxed-log., Exponential	Partial, MRCA age in years (95% HPD interval), Relaxed-log., Exponential
A	8221 (7438, 9076)	7403 (6474, 8543)	6699 (5566, 8148)	6423 (5373, 7719)
B	4883 (4264, 5556)	9754 (8379, 11471)	4590 (3468, 6025)	8599 (6969, 10719)
C	5501 (4782, 6356)	8050 (6442, 9824)	4886 (3499, 6534)	7024 (4564, 10000)
D	4158 (3669, 4683)	4035 (3426, 4813)	3921 (3077, 5044)	3913 (3004, 5040)
E	1016 (712, 1358)	1211, (774, 1756)	857 (331, 1670)	1011 (390, 1892)
G	181 (96, 287)	N/A	173 (70, 334)	N/A
F	4737 (4053, 5538)	3973 (3112, 4966)	4068 (2698, 5679)	3971 (2676, 5595)
H	1713 (1365, 2109)	1232 (819, 1686)	1539 (832, 2498)	1045 (493, 1767)
F/H	7341 (6224, 8603)	6820 (5290, 8560)	6040 (4001, 8637)	5961 (3737, 8641)
Chimp./Gorilla	5281 (4469, 6146)	5355 (4138, 6749)	4523 (2653, 6508)	4669 (2687, 6977)
Gibbon/J/Orang.	7391 (6384, 8484)	8722 (7208, 10590)	5604 (3852, 7700)	6630 (4475, 9189)
RISE154/RISE254 /RISE563	8239 (7502, 9141)	7864 (6769, 9202)	6716 (5626, 8078)	6954 (5692, 8572)

Table 2.7: Median MRCA age of individual nodes under a strict clock and Bayesian skyline population prior or under a relaxed log-normal clock and coalescent exponential population prior. Columns prefaced with 'Partial' were inferred using a partial alignment of positions 1 – 1400 of dataset 2, columns prefaced with 'Complete' were inferred using the complete genome. Abbreviations are as follows: 'Strict': strict clock, 'Relaxed-log.': relaxed log-normal clock, 'Constant': coalescent constant population prior, 'Exponential': coalescent exponential population prior, 'Skyline': coalescent Bayesian skyline population prior.

Model (clock, population prior)	Root age in years (95% HPD interval)	Substitution rate (substitutions/site/year)
Complete, Strict, Constant	16038 (14023, 18140)	$9.27x10^{-6}$ ($8.14x10^{-6}$, $1.05x10^{-5}$)
Partial, Strict, Constant	16385 (13136, 20067)	$8.43x10^{-6}$ ($6.81x10^{-6}$, $1.02x10^{-5}$)
Complete, Strict, Exponential	15840 (13969, 18056)	$9.24x10^{-6}$ ($8.04x10^{-6}$, $1.04x10^{-5}$)
Partial, Strict, Exponential	15501 (12526, 18855)	$8.46x10^{-6}$ ($6.86x10^{-6}$, $1.03x10^{-5}$)
Complete, Strict, Skyline	15593 (13688, 17795)	$9.48x10^{-6}$ ($8.3x10^{-6}$, $1.07x10^{-5}$)
Partial, Strict, Skyline	15347 (12330, 18492)	$8.85x10^{-6}$ ($7.31x10^{-6}$, $1.07x10^{-5}$)
Complete, Relaxed-log. Constant	14561 (9551, 20940)	$1.16x10^{-5}$ ($8.63x10^{-6}$, $1.45x10^{-5}$)
Partial, Relaxed-log. Constant	15086 (9462, 22092)	$1.0x10^{-5}$ ($7.53x10^{-6}$, $1.29x10^{-5}$)
Complete, Relaxed-log. Exponential	11623 (8613, 15275)	$1.18x10^{-5}$ ($9.21x10^{-6}$, $1.45x10^{-5}$)
Partial, Relaxed-log. Exponential	12064 (8263, 16593)	$1.04x10^{-5}$ ($7.68x10^{-6}$, $1.32x10^{-5}$)
Complete, Relaxed-log. Skyline	12923 (8762, 18415)	$1.23x10^{-5}$ ($9.68x10^{-6}$, $1.51x10^{-5}$)
Partial, Relaxed-log. Skyline	13085 (8052, 18184)	$1.07x10^{-5}$ ($8.18x10^{-6}$, $1.34x10^{-5}$)

Table 2.8: Median root age and substitution rates under different clock models and population priors. Abbreviations are as follows: 'Strict': strict clock, 'Relaxed-log.': relaxed log-normal clock, 'Constant': coalescent constant population prior, 'Exponential': coalescent exponential population prior, 'Skyline': coalescent Bayesian skyline population prior. Rows prefaced with 'Partial' were inferred using the genome segment between positions 0 to 1400, rows prefaced with 'Complete' were inferred using the complete genome.

site in Russia, but their viruses have only 93.34% sequence identity (Table 2.12), which indicates the existence of substantial localized HBV diversity about 4.2 ka. The ancient sequence HBV-DA45 phylogenetically groups with genotype B and has 97.65% sequence identity with modern genotype B (Table 2.11). Sequences HBV-DA27, HBV-DA29, HBV-DA51 and HBV-DA222 phylogenetically group with the modern genotype D. They have high sequence identity (96.99–98.74%) with modern genotype D sequences (Table 2.11), and have the typical 33-nucleotide deletion in the preS1 region, encoding the three HBV surface proteins [193] (Table 2.10). Sequences HBV-RISE154, HBV-RISE254, and HBV-RISE563 are in a sister relationship with the chimpanzee-gorilla HBV clade (Fig. 3.6). HBV-RISE254 and HBV-RISE563 have the same 33-nucleotide deletion in the preS1 sequence that is shared with NHP HBVs and human genotype D (Table 2.10). HBV-RISE563 does not encode a functional pre-core peptide (Table 2.10). On the basis of sequence similarity across the whole genome, HBV-RISE563 and HBV-RISE254 together might be classified as a new human HBV genotype that is extinct today, and HBV-RISE154 might possibly be classified as another (Tables 2.11, 2.12). However, HBV-RISE154 has low genome coverage, which precludes an exact calculation. The sister relationship of these three sequences with modern chimpanzee and gorilla HBVs could be interpreted as a consequence of relatively recent transmission(s) of HBV from humans to NHPs [197]. However, other scenarios and confounding factors are possible, as these lineages are

deeply separated in the tree. Incomplete lineage sorting combined with viral extinction (possibly boosted by massive recent reductions in great ape populations) should be considered. More data on current and, if possible, ancient HBVs will be necessary to reach definitive conclusions. The geographical locations of some of the ancient virus genotypes do not match the present-day genotype distribution, and also do not match dates and/or locations inferred in previous studies of HBV. Although the data presented here are limited, they provide important spatio-temporal reference points in the evolutionary history of HBV. Their synopsis suggests a more complicated ancestry of present-day genotypes than previously assumed, especially in light of recent insights into the history of human migration.

We find genotype A in south-western Russia by 4.3 kya (in samples RISE386 and RISE387) in individuals belonging to the Sintashta culture, and in a Hungarian sample (DA195) from the Scythian culture. The western Scythians are related to the Bronze Age cultures of western steppe populations [44] and their shared ancestry suggests that the modern genotype A may descend from this ancient Eurasian diversity and not, as previously hypothesized, from African ancestors [238, 239]. This is also consistent with the phylogeny (Fig. 3.6), as well as the fact that the three oldest ancient genotype A sequences (HBV-DA195, HBV-RISE386 and HBV-RISE387) lack the six-nucleotide insertion found in the youngest (HBV-DA119) and in all modern genotype A sequences. The ancestors of subgenotypes A1 and A3 could have been carried into Africa subsequently, via migration from western Eurasia [240]. The ancient HBV genotype D sequences were all found in Central Asia. HBV-DA27, found in Kazakhstan and dated to 1.6 ka, falls basal to the modern subgenotype D5 sequences that today are found in eastern India [241]. Based on the observation that genotypes go extinct and can be created by recombination, the ancient sequence data show that the diversity that we observe today is only a subset of the diversity that has ever existed. These data support a scenario in which all present-day HBV diversity arose only after the split of the Old World and New World genotypes (25–13.4 ka). Any attempt to interpret the currently known HBV tree based on human migrations that happened before this event will necessarily result in anomalies that cannot be reconciled, such as the basal position of genotypes F and H and the apical position of subgenotype C4, which is found exclusively in indigenous Australians. If HBV did co-evolve with ancient modern humans as they left Africa as previously proposed [191], most of the pattern of earlier diversity has been replaced by changes that happened after the split of the Old and New World genotypes. Genotypes F and H would therefore be remnants of the earlier now-extinct diversity, and the arrival of subgenotype C4 in Australia would have taken place long after the split between Old and New World genotypes, as supported by the tree in Fig. 3.6. Alternatively,

there could have been a New World origin of HBV or the virus could have been introduced into humans from a different host. Our data do not allow us to speculate either way.

To our knowledge, we report the oldest exogenous viral sequences recovered from DNA of humans or any vertebrate, and show that it is possible to recover viral sequences from samples of this age. We show that humans throughout Eurasia were widely infected with HBV for thousands of years. Despite the age of the samples and the imperfect diagnostic test, our dataset contained a high proportion of HBV-positive individuals. The actual ancient prevalence during the Bronze Age and thereafter might have been higher, reaching or exceeding the prevalence typically found in contemporary indigenous populations [188].

This clearly establishes the potential of HBV as powerful proxy tool for research into human spread and interactions. The data from ancient genomes reveal aspects of complexity in HBV evolution that are not apparent when only modern sequences are considered. They show the existence of ancient HBV genotypes in locations incongruent with their present-day distribution, contradicting previously suggested geographical or temporal origins of genotypes or sub-genotypes; evidence for the creation of genotype A via recombination and the emergence of the genotype outside Africa; at least one now-extinct human genotype; ancient genotype-level localized diversity; and provide further evidence that the magnitude of viral substitution rates inferred from heterochronously sampled sequences is dependent on the time interval over which the sequences were sampled. Together, these findings suggest that the difficulty in formulating a coherent theory for the origin and spread of HBV may be due to genetic evidence of an earlier evolutionary scenario being overwritten by relatively recent alterations, as has previously been suggested in the context of recombination [235]. The lack of ancient sequences limits our understanding of the evolution of HBV and very probably of other viruses. Discovery of additional ancient viral sequences may provide a clearer picture of the true origin and early diversification of HBV, enable us to address questions of palaeo-epidemiology, and broaden our understanding of the contributions of natural and cultural changes (including migrations and medical practices) to human disease burden and mortality.

Hypothesis	Strict, Skyline			Relaxed-log.,	Exponential	
	Node age (years)	Likelihood	Bayes factor	Node age	Likelihood	Bayes factor
A3 (200: 500)	841	-50502.94	3.3	667	-50388.15	63.36
C3 (5100: 12,000)	1897	-50532.75	33.12	1983	-50388.13	63.34
F/H (13,400: 25,000)	7341	-50493.15	-6.48	6040	-50388.05	63.25

(a) Path sampling results computed using the complete alignment of dataset 2.

Hypothesis	Strict, Skyline			Relaxed-log.,	Exponential	
	Node age (years)	Likelihood	Bayes factor	Node age	Likelihood	Bayes factor
A3 (200: 500)	538	-19419.2	-3.85	456	-19383.79	-1.34
C3 (5100: 12,000)	1,844	-19443.99	20.94	1,745	-19384.39	-0.73
F/H (13,400: 25,000)	6,820	-19392.12	-30.93	5,961	-19355.51	-29.61

(b) Path sampling results computed using the partial alignment (positions 1 – 1400) of dataset 2.

Table 2.9: Results of testing different calibration point hypotheses under a strict clock and Bayesian skyline population prior or under a relaxed log-normal clock and coalescent exponential population prior. The 'Hypothesis' column refers to the following hypothesis tests: A3: the split of the Haitian A3 sequences (FJ692611, FJ692598) took place 200 - 500 years ago. C3: the split of the Melanesian C3 sequences (X75656, X75665) took place 5,100 - 12,000 years ago. F/H: the split of genotypes F and H took place 13,400 - 25,000 years ago. Nodes were constrained using an uniform distribution. Positive values for the Bayes factor indicate support for the model without the constraint, negative values support the model with the constraint for the hypothesis that was tested. The column 'Node age (years)' shows the age of the node to which the hypothesis test applies when it is inferred without constraints. Abbreviations are as follows: 'Strict': strict clock, 'Relaxed-log.': relaxed log-normal clock, 'Exponential': coalescent exponential population prior, 'Skyline': coalescent Bayesian skyline population prior, 'Likelihood': log marginal likelihood estimate.

Sample	Genotype of closest modern sequence	Sequence identity to closest modern sequence	Genome length	Predicted serotype	Insertions / deletions	Predicted HBeAg status
DA119	A3	97.8%	3221	*adw2*	6-nucleotide insertion at the C terminus of the core region	Positive
DA195	A3	96.1%	3215	*adw2*	None	Positive
RISE386	A	95.3%	3215	*adw2*	None	Positive
RISE387	A	92.5%	3215	*adw2*	None	Negative; PreC stop codon
DA45	B1	97.6%	3215	*ayw1*	None	Positive
DA29	D3	98.6%	3182	*ayw2*	33-nucleotide deletion at the N terminus of the preS1 region	Positive
DA222	D3	98.7%	3182	*ayw2*	33-nucleotide deletion at the N terminus of the preS1 region	Positive
DA27	D5	97.2%	3182	*ayw2*	33-nucleotide deletion at the N terminus of the preS1 region	Positive
DA51	D1	97%	3182	*ayw2*	33-nucleotide deletion at the N terminus of the preS1 region	Positive
RISE154	Chimpanzee	92.5%	Ambiguous	*adw2**	Ambiguous	Positive
RISE254	Chimpanzee	94%	3182	*adw2*	33-nucleotide deletion at the N terminus of the preS1 region	Positive
RISE563	Gorilla	92.7%	3182	*adw2*	33-nucleotide deletion at the N terminus of the preS1 region	Negative; PreC stop codon

Table 2.10: Genome properties of ancient sequences included in phylogenetic analyses. Genotype groups are sorted by increasing sample age. The serotype was predicted following Table 1 in Kay and Zoulim, 2007 [242]. *Serotype could not be determined unambiguously, owing to lack of coverage.

Sample	A	B	C	D	E	F	G	H	I	J	Chimp	Gibbon	Gorilla	Orang Utan
DA119	**97.27**	92.60	92.94	92.38	92.05	88.42	90.62	88.24	93.32	89.78	92.10	91.15	92.95	90.98
DA195	**96.11**	92.16	92.59	92.06	92.00	88.24	90.32	87.68	93.22	89.37	92.13	91.65	92.32	90.99
RISE386	**95.32**	92.55	92.96	92.83	91.87	88.57	90.22	87.94	93.79	89.60	92.33	91.43	92.64	91.34
RISE387	**92.49**	90.70	91.70	91.74	91.06	88.57	91.42	88.60	91.54	88.96	91.55	90.75	91.37	90.65
DA45	91.89	**97.65**	96.04	90.47	90.55	87.70	88.11	88.06	91.07	89.40	91.50	91.79	91.58	90.53
DA27	91.67	91.28	92.96	**97.21**	93.15	88.77	89.93	87.83	91.08	89.63	92.43	92.24	92.57	90.31
DA29	91.46	90.50	92.79	**98.64**	93.21	87.82	89.41	87.64	90.82	88.66	91.10	90.74	91.94	89.66
DA51	92.10	91.32	93.24	**96.99**	93.47	88.48	90.28	88.04	91.11	89.30	91.55	91.54	92.26	90.40
DA222	92.29	91.61	93.88	**98.74**	93.55	88.66	90.01	88.23	91.45	89.23	92.16	91.99	93.29	90.63
RISE154	90.86	91.07	92.11	91.76	91.81	89.20	91.26	88.91	91.62	90.05	**92.54**	92.36	92.26	92.36
RISE254	92.04	91.66	92.32	92.56	92.80	89.25	91.20	88.86	91.49	90.47	**94.03**	93.25	93.93	92.96
RISE563	90.63	90.43	90.70	91.41	91.27	88.83	90.51	88.75	90.27	89.16	92.30	91.94	**92.66**	91.08

Table 2.11: Best consensus sequence identity with 14 groups of HBV full genomes. The Needleman-Wunsch algorithm (as implemented in EMBOSS50) was used to globally align each sample consensus sequence against each of the 3,384 full HBV genomes of Dataset 4 (see Methods). The table shows the best nucleotide similarity percentage for each sample consensus against 14 genome groups from the full set of HBV genomes. In cases in which the consensus length is less than the genome length, the given figure is the percentage of identical nucleotides in the matching region, not counting any alignment gaps or ambiguous consensus nucleotides. For each sample, the genome group with the highest identity is highlighted in bold. *Table contributed by Terry Jones.*

Sample	DA195	RISE386	RISE387	DA45	DA27	DA29	DA51	DA222	RISE154	RISE254	RISE563
DA119	96.79	95.99	92.69	92.08	91.87	91.97	92.37	92.41	91.39	92.23	90.78
DA195		95.86	92.72	91.83	92.15	91.50	92.47	91.83	91.34	91.91	90.36
RISE386			93.34	92.34	92.74	92.55	92.97	92.64	91.66	92.63	91.07
RISE387				90.60	91.50	91.33	91.89	91.75	91.78	92.12	91.26
DA45					89.73	89.39	90.75	90.38	90.43	90.97	89.66
DA27						96.42	97.24	97.31	91.45	92.49	91.39
DA29							96.62	98.06	90.72	91.52	90.48
DA51								97.12	91.83	92.66	91.44
DA222									91.57	92.53	91.28
RISE154										94.39	93.02
RISE254											95.77

Table 2.12: Inter-consensus sequence identity. The Needleman-Wunsch algorithm was used to globally align all ancient sample consensus sequences against one another. The table shows the nucleotide identity percentage for each alignment. In cases in which the consensus lengths were unequal, the given figure is the percentage of identical nucleotides in the matching region, not counting any alignment gaps or ambiguous consensus nucleotides. *Table contributed by Terry Jones.*

CHAPTER 3: ANCIENT HUMAN PARVOVIRUS B19 IN EURASIA REVEALS ITS LONG-TERM ASSOCIATION WITH HUMANS

PREFACE

A version of this chapter was previously published as (* denotes equal contribution):

Barbara Mühlemann*, Ashot Margaryan*, Peter de Barros Damgaard*, Morten E. Allentoft*, Lasse Vinner, Anders J. Hansen, Andrzej Weber, Vladimir I. Bazaliiskii, Martyna Molak, Jette Arneborg, Wieslaw Bogdanowicz, Ceri Falys, Mikhail Sablin, Vàclav Smrčka, Sabine Sten, Kadicha Tashbaeva, Niels Lynnerup, Martin Sikora, Derek J. Smith, Ron A. M. Fouchier, Christian Drosten, Karl-Göran Sjögren, Kristian Kristiansen, Eske Willerslev, Terry C. Jones. *Ancient human parvovirus B19 in Eurasia reveals its long-term association with humans.* PNAS, 115(29), 7557–7562, (2018).

It has been modified to fit the style of a dissertation.

I did all computational analyses, figures and tables, except the recombination analysis, and figures and tables related to that, which were done by Terry Jones. The sequencing work was performed by Ashot Margaryan, Peter de Barros Damgaard and Morten Allentoft. Targeted virus capture of the DA251 sample was performed by Lasse Vinner. I wrote the text with input from Terry Jones and the other co-authors. Andrzej Weber, Jette Arneborg, Niels Lynnerup, Ceri Falys, Wieslaw Bogdanowicz, and Martyna Molak-Tomsia wrote the section on the archaeological context of samples in the Supplementary Information. In addition to this description, I have noted in the legends of figures and tables if they were contributed by others.

3.1 ABSTRACT

Human parvovirus B19 (B19V) is a ubiquitous human pathogen associated with a number of conditions, such as fifth disease in children and arthritis and arthralgias in adults. B19V is thought to evolve exceptionally rapidly among DNA viruses, with substitution rates previously estimated to be closer to those typical of RNA viruses. On the basis of genetic sequences up to ~70 years of age, the most recent common ancestor of all B19V has been dated to the early 1800s, and it has been suggested that genotype 1, the most common B19V genotype, only started circulating in the 1960s. Here we present 10 genomes (63.9–99.7% genome coverage) of B19V from dental and skeletal remains of individuals who lived in Eurasia and Greenland from ~0.5 to ~6.9 thousand years ago (kya). In a phylogenetic analysis, five of the ancient B19V sequences fall within or basal to the modern genotype 1, and five fall basal to genotype 2, showing a long-term association of B19V with humans. The most recent common ancestor of all B19V is placed ~12.6 kya, and we find a substitution rate that is an order of magnitude lower than inferred previously. Further, we are able to date the recombination event between genotypes 1 and 3 that formed genotype 2 to ~5.0 to ~6.8 kya. This study emphasizes the importance of ancient viral sequences for our understanding of virus evolution and phylogenetics.

3.2 INTRODUCTION

Infection with human parvovirus B19 (B19V) can have a number of different outcomes, from asymptomatic or non-specific symptoms to fifth disease (erythema infectiosum) in children, arthritis and arthralgias in adults, and hydrops fetalis in pregnant women. It can also lead to transient or persistent erythroid aplasia and aplastic crisis in people with underlying hematological disorders [243, 244]. Management of B19V infections is generally limited to symptomatic treatment, as there are currently no vaccines or antivirals [243]. The virus replicates in erythroid precursor cells in the bone marrow [245]. After the initial infection, viral DNA is detectable in multiple human tissues, including kidney, lymph nodes, heart, testes, and synovial tissue, in symptomatic and asymptomatic individuals [246–251].

The persistence of viral DNA in an infected individual is thought to be lifelong, and the persisting virus corresponds to the virus genotype of the initial infection [252]. B19V can be transmitted via the respiratory or blood-borne route. The virus can be detected in the saliva and is present in the blood in high concentrations (10^{13} virions/ml) during the viraemic phase of the infection [243, 253]. In blood donors, seroprevalence of B19V is around 60% on average, and the rate of new individuals infected per year is estimated at between 0.5% and 1% [254].

B19V is a single-stranded DNA (ssDNA) virus with a genome of ~5,600 nucleotides. The coding region is flanked by terminal repeats of 383 nucleotides and contains three major proteins: the nonstructural protein NS, the capsid proteins VP1 and VP2, and the minor proteins 11 kDa, 9 kDa, and 7.5 kDa [243].

B19V belongs to the *Erythrovirus* genus in the *Parvoviridae* family [243]. There are three genotypes, with genotype 1 having ~10% sequence divergence from genotypes 2 and 3 and genotypes 2 and 3 having ~5% sequence divergence between each other [255]. Genotypes 1 and 3 are further split into two subgroups, a and b, with sequence divergence of about 5% [256, 257]. All three genotypes form a single serotype [258]. The distribution of the three genotypes is not spatially and temporally uniform: genotype 1 has a worldwide distribution [259], genotype 2 is found mainly in elderly adults in northern Europe [252], and genotype 3 is found in Sub-Saharan and West Africa, South America, and France [256]. Because genotype 2 is today only found in tissues of people born before the 1970s, it has been hypothesized that it was replaced by genotype 1 and that genotype 1 originated in the second half of the 20^{th} century [158].

B19V is deemed unique among the DNA viruses because its substitution rate, inferred from modern heterochronous sequences, is unusually high, in the range of 1.0–4.0x10^{-4} nucleotide substitutions per site per year (s/s/y) [126, 157, 158], which is more similar to substitution rates of RNA viruses than either single-stranded or double-stranded DNA viruses [35].

Recent advances in sequencing ancient DNA (aDNA) have provided important insights into past human population dynamics [40] and the evolution of bacterial human pathogens [78, 105], but have only recently been applied to viruses [79, 116]. Viral sequences recovered from aDNA samples have provided important reference points for the calibration of molecular clocks [79, 116]. These samples can also provide insight into the spatiotemporal distribution of past viral variants and strains [79]. B19V has characteristics that should favor molecular preservation, including a DNA genome, high viraemia, and a stable nonenveloped virion [260]. Indeed, the persistence of B19V DNA in bones from individuals deceased ~70 years ago has been established [126]. Here, we extend our knowledge of the virus further back in time and present 10 B19V coding region sequences (63.9 to 99.7% genome coverage) recovered from humans living in Eurasia and Greenland between ~0.5 and 6.9 kya, leading to an improved understanding of the timescale of B19V evolution.

3.3 METHODS

3.3.1 Archaeological context of samples

Information about the archaeological context of samples was contributed by Andrzej Weber, Jette Arneborg, Niels Lynnerup, Ceri Falys, Wieslaw Bogdanowicz, and Martyna Molak-Tomsia, and can be found in the Supplementary Information to [80].

3.3.2 Sample preparation, capture and sequencing

Sample preparation and sequencing were performed as described in Allentoft *et al.,* (2015) [41]. One sample (DA251) was selected for virus capture. Capture was performed as described in Mühlemann *et al.,* (2018) [79] with pre-treatment by uracil-specific excision reagent (USER).

3.3.3 Datasets

Below follows a description of the datasets used in this study:

Dataset 1:
Used for bwa and blast analysis, 11 complete genomes:
Genotype 1: FN669502, DQ357065, AF113323
Genotype 2: AJ717293, DQ333427, HQ340602
Genotype 3: NC004295, AY083234, AY083234, AJ249437
Bovine parvovirus: NC_001540

Dataset 2:
Reference sequences used to make consensuses in Geneious. Only the coding region was used.
Genotype 1: M13178
Genotype 2: AY044266
Genotype 3: AJ249437

Dataset 3:
Representative dataset of complete parvovirus B19 genomes used for genotype assignment. All sequences under taxid 10798 (Human parvovirus B19) and taxid 344889 (unclassified Erythrovirus) were downloaded from NCBI on 18 October 2017.

Sequences that were shorter than 4000 nucleotides were removed, as well as sequences from patents or sequences designated as artificial or recombinant. Only coding region was used. The final dataset contains 102 sequences:
KX752821, AJ781038, KM393165, KM393168, KM393167, KM393166, AB126268, AB126267, AB126266, KM065415, KC013340, KT310174, KR005643, DQ225150, DQ225149, KR005644, KR005641, KR005640, KR005642, NC_000883, AY386330, KM065414, KM393163, M13178, FN598218, FN598217, Z70560, Z68146, DQ4083-01, AB030693, AB030673, KM393-164, AY028237, KC013329, KC013325, Z70528, KT268312, KC013305, AF113323, M24682, Z70599, AY504945, AB126271, AB1-26264, AB126263, AB126262, AB126269, AF162273, FJ591158, KC013343, KC0-13324, KC013308, KM393169, AB030694, DQ293995, KC013316, DQ225151, KC013321, KC013344, KC013338, KC013331, KC013314, KC013312, KC013333, KC013346, KC013351, KC013313, KC013327, KC013332, AB126265, KC013322, DQ225148, AB126270, DQ357065, DQ357064, KF724387, AY903437, AY044266, DQ333426, AB550331, EF216869, AJ717293, KF724386, DQ333428, AY064476, AY064475, AY647977, AY083234, AY582124, DQ234779, DQ234778, DQ408305, DQ408302, DQ408304, DQ408303, FJ265736, AY582125, DQ234775, DQ234771, DQ234769, NC_004295, AJ249437.

Dataset 4:
Dataset used for phylogenetic and saturation analyses. The dataset contains 77 sequences from dataset 3 in addition to the 10 ancient sequences analysed in more detail:
AJ781038, KM393165, KM393168, KM393166, AB126267, AB126266, KM065415, KC013340, KT310174, DQ225150, DQ225149, KR005641, KR005640, AY386330, KM393163, M13178, FN598218, Z70560, Z68146, DQ408301, AB030673, KM393-164, KC013329, KC013325, KT268312, KC013305, AF113323, Z70599, AY504945, AB126271, AB126262, AB126269, AF162273, FJ591158, KC013343, KC013324, KC013308, KM393169, AB030694, DQ293995, KC013316, DQ225151, KC013321, KC013312, KC013333, KC013346, KC013313, KC013327, AB126265, AB126270, DQ357065, DQ357064, KF724387, AY903437, AY044266, DQ333426, AB550331, EF216869, AJ717293, KF724386, DQ333428, AY064476, AY064475, AY647977, AY083234, AY582124, DQ234779, DQ234778, DQ408305, DQ408302, DQ408304, DQ408303, FJ265736, DQ234775, DQ234771, DQ234769, AJ249437, DA251, DA336, DA337, RISE569, NEO105, VK6, VK143, VK154, VK477, DA66.

3.3.4 Identification, consensus sequence generation, and authentication of ancient Parvovirus B19

We screened shotgun sequencing data from 1578 individuals from Eurasia from 0.2 to 24 kya for the presence of reads matching B19V. AdapterRemoval [205] (version 2.1.7) was used with its default settings to remove adaptors from all sequences, to trim N bases from the ends of reads, and to trim bases with quality <= 2. Reads were aligned against a human genome (GRCh3855) using bwa [206] (version 0.7.15-r1140, mem algorithm). Reads that did not match the human genome were aligned to Dataset 1 using bwa (version 0.7.15-r1140, mem, aln, and aln `-l` algorithms). Samples for which the reads matching B19V covered >30% of the coding region of the B19V genome were also aligned to Dataset 1 using BLASTn [25] (version 2.4.0) (with arguments `-task blastn -evalue 0.01`). One sample (DA251) was selected for virus capture (see above). The resulting reads were also aligned to Dataset 1 using BLASTn, as described above, and matching reads were added to the reads matching B19V using bwa and BLASTn from DA251. Reads matching using bwa and BLASTn were combined and compared against the complete NCBI nt database downloaded on 28-12-2016. Only reads with their highest match against sequences that contained any of the words 'B19', 'Parvo', or 'Erythro' (case insensitive), were retained. To make the consensus sequences, reads were aligned to Dataset 2 in Geneious [211] (version 9) using Medium sensitivity / fast and iterate up to 5 times. The reference with the highest number of matching reads was used as a template for constructing the consensus sequence. Alignments were inspected and damaged ends were trimmed manually. Reads were checked for $C \rightarrow T$ damage at the 5' end, typical for aDNA using mapDamage [209] (version 2.0.6).

3.3.5 Phylogenetic analysis

All phylogenetic analyses were performed using a Mafft alignment [221] (version 7) of Dataset 4. A maximum likelihood phylogeny was inferred using PhyML [222] (version 20160116). The TN93 substitution model was used and topology, branch lengths, and rates were optimized. 100 bootstrap replicates were performed. The final trees show nodes with support values less than 70 as polytomies. Saturation was assessed using DAMBE [261] (version 6). Transition and transversion frequencies were plotted using the graphics option. Formal saturation tests were performed in DAMBE using the method of Xia [261, 262], using a proportion of 0.3 invariant sites. Root-to-tip regressions were performed on three different trees: 1) A neighbor joining tree inferred in Geneious using the TN93 substitution model. 1000 bootstrap

replicates were performed. 2) The ML tree inferred above. 3) A Bayesian tree inferred in MrBayes [224] (version 3.2.5), using the TN93 substitution model, invariant sites, and gamma distributed rate heterogeneity among sites. Root-to-tip distances were extracted in TempEST [225] (version 1.5), using the 'best-fitting-root' parameter, and the regression analysis was performed in SciPy [226]. Date randomization tests were performed in BEAST2 [150] (version 2.4.4, prerelease) to test whether the temporal signal in the data is sufficient for a dated phylogenetic analysis. Using bModelTest [227] (version 1.0.4), we selected a 123,143 substitution model with unequal base frequencies, four gamma rate categories, estimated gamma distribution of rate variation, and estimated proportion of invariant sites. We used a strict clock and coalescent constant population prior. The clock rate was constrained using a uniform ($1x10^{-9}$ to $1x10^{-3}$ s/s/y) prior. Two different randomizations were performed in replicates of five: 1) Tip dates were randomized for all sequences. 2) Only tip dates of the ancient sequences were randomized. As a criterion for a temporal signal, we employed CR2 described in [228], where there is evidence for a temporal signal if none of the 95% HPD intervals of the substitution rate overlap between the randomized and the correctly dated run. The Markov chain Monte Carlo (MCMC) analysis was run for 40,000,000 generations, all parameters reached an effective sample size (ESS) >200, sampling every 2000 generations. Dated coalescent phylogenies were inferred using BEAST2 [150]. We used a 123,143 substitution model with unequal base frequencies, four gamma rate categories, estimated gamma distribution of rate variation, and estimated proportion of invariant sites. The molecular clock was calibrated using tip dates. Proper priors were used throughout; specifically, we used a uniform ($1x10^{-9}$ to $1x10^{-3}$ s/s/y) prior on the clock rate. Path sampling, as implemented in BEAST2, was used to select the best fitting population prior: per path sampling run, 50 steps with a chain length of 1,000,000 generations were run (Table 3.5). Log marginal likelihood values were compared using a Bayes factor test. For the final tree shown in Figure 3.6, a strict clock and coalescent Bayesian skyline population prior were used. The MCMC analysis was run for 60,000,000 generations until all parameters reached an ESS >200, sampling every 2000 generations. Furthermore, a tree was inferred under a relaxed log-normal clock and coalescent Bayesian skyline population prior. Genotypes were constrained to be monophyletic, as supported by the maximum likelihood tree (Fig. 3.5). The MCMC chain was run for 226,000,000 generations, sampling every 2000 generations. Convergence and mixing were assessed using Tracer [230] (version 1.6). Additionally, trees were inferred for each genotype separately, for the minor (positions 1877–3352) and major (positions 1–1876, 3353–4354) parent sections of the genome, and for genotype 1 and 3 sequences together using the same parameters as for the final tree above. The MCMC

chain was run for 40,000,000 generations, except for the genotype 2 separately, which was run for 60,000,000 generations. We sampled every 2000 generations. The final tree files were subsampled to contain 20,000 trees, or 22,500 for the strict clock and coalescent Bayesian skyline run, with the first 25% of samples discarded as burn-in. Maximum clade credibility trees were calculated using TreeAnnotator (version 2.4.4 prerelease).

3.3.6 Genotype assignment

Percent sequence identity was calculated between all sequences in dataset 3. Locations with undefined bases were omitted.

3.3.7 Recombination analysis

The recombination analysis described in this section was performed by Terry Jones. Recombination analysis was performed using seven algorithms built into RDP4 [214]. The algorithms are RDP [263], GENECONV [215], BootScan [216], MaxChi [217], Chimaera [218], SiScan [219], and 3Seq [220]. The analysis involved the 10 ancient sequences of this paper, plus the 13 following modern sequences:

Genotype 1: AB126262, AB126267, AB126270, DQ357064, DQ357065, M13178
Genotype 2: AY044266, DQ333426, KF724386
Genotype 3: AY582124, DQ234775, DQ408305, FJ265736

3.4 RESULTS

3.4.1 Identification and authentication

We screened shotgun DNA sequencing data representing dental or skeletal remains of 1,578 ancient human individuals (~0.2 to 24.0 kya) recovered from across Eurasia, Southeast Asia, and Greenland for the presence of reads matching B19V. A total of 20 samples had reads covering >30% of the coding region of B19V, 10 of which had coverage >50% (Table 3.2). The samples with coverage >50% were between ~0.5 and ~6.9 thousand years old and from different archaeologically defined cultures (four Viking Age Scandinavians, three Early Neolithic to Early Bronze Age Baikal Hunter-Gatherers, one early Slav from the Czech Republic, and one Tian Shan Hun) (Table 3.2). The individuals came from a wide geographic range, spanning Europe, Central Asia, and Greenland (Fig. 3.1 and Table 3.2). The samples with coverage >50% showed DNA damage patterns typical for aDNA [67] when at least 100 reads were available (8 of 10 samples; Fig. 3.2). Clear damage patterns could not be definitively identified when there were fewer than 100 reads (2 of 10 samples); there was no evidence for this being for any reason other than a low number of reads. Authenticity of the ancient B19V sequences presented here is further supported by our compliance with standard precautions for working with aDNA [63] and because B19V sequences were found in only 20 of 1,578 ancient human individuals, which would not be expected in case of ubiquitous laboratory contaminant. Finally, the ancient sequences mostly occupy basal positions in the phylogenetic tree, with short branch lengths from the trunk, for which the only explanation is that they are ancestors of modern strains.

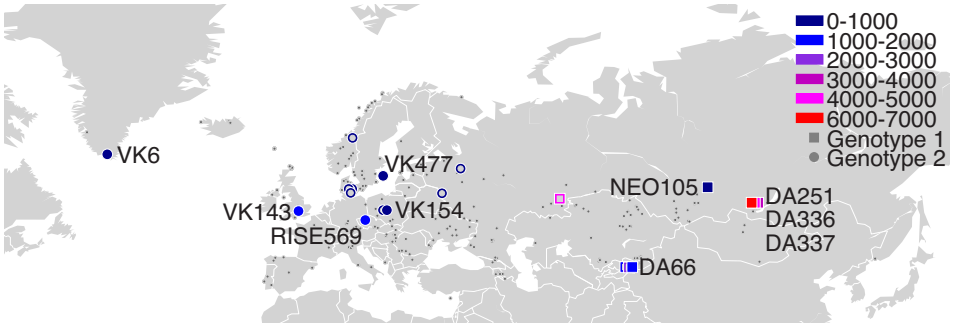


Figure 3.1: Geographic locations of the samples with reads matching B19V. Samples are coloured by age. Squares indicate genotype 1, circles genotype 2. Empty symbols indicate samples with 30–50% coverage of the coding region, filled symbols samples with >50% coverage of the coding region that were included for further analysis. Samples that were negative for B19V are shown as small grey symbols (samples outside Eurasia are omitted for clarity).

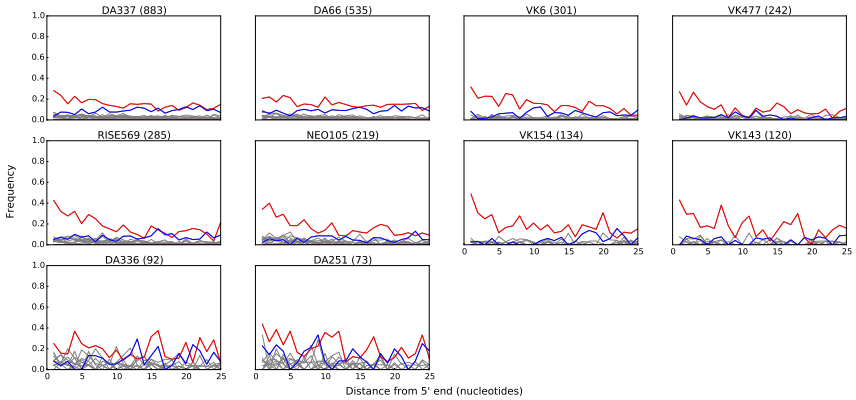


Figure 3.2: Ancient DNA damage patterns for the samples included for further analysis. The frequencies of the mismatches observed between the B19V reference sequences (see Table 3.1) and the reads are shown as a function of distance from the 5' end. C $\rightarrow$ T (5') and G $\rightarrow$ A (3') mutations are shown in red and blue, respectively. All other possible mismatches are reported in grey. Insertions are shown in purple, deletions in green, and clippings in orange. The number of reads matching B19V for each sample is shown in parentheses. Note that the number of reads for DA251 is lower here than in Table 3.1. This is because Table 3.1 includes the reads from the USER treated sequencing (see Methods).

	BWA							BLASTn		BWA +	BLASTn				
Sample	Reads (bwa aln -l)	Reads (bwa aln)	Reads (bwa mem)	Total deduplicated reads	Reads aligned in Geneious	Coverage of consensus	Coverage depth	Reads (blastn)	Bit score cutoff	Total reads	highest match to parvo	Reads aligned in Geneious	Coverage of consensus	Coverage depth	Reference*
VK139	255	226	494	477	38	40%	0.5x	799	40	799	617	46	40.8%	0.5x	2
VK141	315	290	724	559	39	33.3%	0.3x	1210	40	1210	701	41	35%	0.5x	2
VK143	463	428	991	747	109	73.2%	1.6x	1144	40	1144	987	120	67.3%	1.6x	2
VK154	359	315	913	711	124	72.6%	1.5x	1771	40	1771	931	134	71.4%	1.6x	2
VK155	564	525	862	242	53	41%	0.7x	347	40	347	273	57	41.5%	0.7x	2
VK160	379	325	669	575	42	74%	0.7x	1086	40	1086	669	57	51.1%	0.8x	2
VK223	164	151	208	197	48	36.6%	0.5x	295	40	295	220	56	36.5%	0.5x	2
VK275	1861	1692	3276	961	59	38.4%	0.7x	1361	40	1361	1219	58	38%	0.7x	2
VK477	2305	2181	3014	2665	238	92.8%	3.2x	3217	40	3219	3027	242	87.7%	3.1x	2
VK548	394	376	514	366	40	42.1%	0.5x	1325	40	1325	1172	44	41.2%	0.5x	2
VK6	2251	2028	3696	3437	284	48.4%	4.0x	4686	40	4686	4675	301	82.5%	4.0x	2
DA55	62	64	101	109	35	33.7%	0.5x	346	40	346	173	63	43.8%	0.8x	1
DA66	541	513	356	692	324	95.7%	4.7x	1123	40	1125	1084	535	99.7%	7.8x	1
DA68	187	183	136	116	64	36.4%	0.5x	185	40	189	161	88	45.7%	0.7x	1
RISE392	107	96	392	151	39	33.2%	0.7x	206	50	207	207	56	36.8%	0.9x	1
RISE569	1494	1366	2304	2307	229	85.5%	2.7x	3238	40	3240	3234	285	84.1%	3.4x	2
DA251	150	140	352	351	45	47.5%	0.7x	663	40	663	458	73	54.2%	1.0x	1
DA251-user								2165	40	N/D	2586	228	82.8%	3.2x	1
DA336	218	198	495	421	74	62.6%	1.1x	762	40	762	498	92	63.9%	1.3x	1
DA337	1378	1267	2649	2480	728	96.6%	11.3x	3369	40	3371	2729	883	98.4%	13.5x	1
NEO105	3967	3768	4623	4109	184	84.1%	2.2x	1838	40	4734	4386	219	83.6%	2.4x	1

Table 3.1: Mapping statistics of samples with reads matching B19V. * Reference accession number: 1: M13178, 2: AY044266.

Sample	Location	Culture	14C age (s.d.)	Mean calBP age (s.d.)	Estimated age (years, AD)	Age (years)	Sex	Individual age	Sample type
VK139	Fyn, DEN	Viking	N/D	N/D	1000	1000	M		tooth
VK141	Fyn, DEN	Viking	N/D	N/D	1000	1000	F		tooth
VK143	Oxford, UK	Viking	N/D	N/D	880–1000	1077	M		petrous
VK154	Bodzia, POL	Viking	N/D	N/D	980/990–1030/1035	1000	F	early adultus	tooth
VK155	Bodzia, POL	Viking	N/D	N/D	1000	1000	F		tooth
VK160	Kurevanikka, RUS	Viking	N/D	N/D	1000	1000	M		tooth
VK223	Gnezdovo, RUS	Viking	N/D	N/D	1000	1000	M		tooth
VK275	Fyn, DEN	Viking	N/D	N/D	1000	1000	M		tooth
VK477	Gotland, SWE	Viking	N/D	N/D	1000	1000	F		tooth
VK548	N. Trondelag, NOR	Viking	N/D	N/D	1000	1000	F		tooth
VK6	Eastern Settlement, GRL	Viking	N/D	N/D	1000	1000	F	>40y.	tooth
DA55	Tian Shan, KGZ	Saka	2034 (32)	1992 (48)	N/A	2059	M		
DA66	Tian Shan, KGZ	Hun	1546 (33)	1451 (47)	N/A	1518	F		
DA68	Tian Shan, KGZ	Hun	1479 (31)	1365 (35)	N/A	1432	F		
RISE392	Stepnoe VII, RUS	Sintashta	3626 (33)	3942 (53)	N/A	4009	M		tooth
RISE569	Brandysek, CZE	Early slav	1300 (30)	1237 (34)	N/A	1304		Inf II	tooth
DA251	Lake Baikal (Shamanka II), RUS	Early neolithic	5955 (72)	6795 (90)	N/A	6862	M	25–35y.	tooth
DA336	Lake Baikal (Shamanka II), RUS	EBA	3723 (70)	4079 (107)	N/A	4146	M		
DA337	Lake Baikal (Shamanka II), RUS	EBA	3593 (67)	3900 (99)	N/A	3967	M?	17–18y.	
NEO105	Bazaiha, RUS	Historical period	415 (25)	477 (42)	N/A	544			tooth, bone

Table 3.2: Overview of samples with reads matching B19V. Columns 4–7 refer to the 14C age in years before the present (BP) and standard deviation, the mean calibrated age BP and standard deviation, the archaeological age estimated from the sample context, and the sample age used in the phylogenetic analyses corresponding to the mean calibrated age BP relative to 2017, respectively.

3.4.2 Similarity to modern genotypes

We compared the ancient sequences with all published non-artificial, complete B19V genomes in GenBank (Fig. 3.5a). Five ancient sequences are most similar to genotype 1 (DA66, DA251, DA336, DA337, NEO105), and five to genotype 2 (RISE569, VK6, VK143, VK154, VK477). The mean pairwise sequence identity between the ancient and the modern sequences is below the mean within-genotype pairwise sequence identity of the modern sequences (blue and red dashed lines in Fig. 3.5a). None of the ancient sequences are more diverged from any modern sequence than the modern genotypes are from each other (dashed black lines in Fig. 3.5a); hence, we consider them part of the current genotypes and do not propose that they should

be classified as new genotypes. We did not find any differences, such as insertions or deletions, between the ancient and the modern sequences. Of the samples with a genome coverage between 30% and 50% that were not included in the phylogenetic analysis, three had the highest number of reads matching genotype 1, and seven best matched genotype 2 (Fig. 3.5b).

3.4.3 Recombination analysis

The write-up, figures (Figs. 3.3 and 3.4) and table (Table 4.3) in this section (section 3.4.3) were contributed by Terry Jones and are included here for completeness.
We performed recombination analyses using RDP4 [214] on the 10 ancient sequences with coverage >50% and a selection of 13 modern genotype 1–3 sequences, selected to represent the modern diversity. All seven recombination detection algorithms used by RDP4 identified all eight genotype 2 sequences included in the analysis (three modern and five ancient sequences) as potential recombinants, with very high probabilities (Table 4.3). Recombination plots for the RDP and MaxChi algorithms of the RDP4 package for all eight genotype 2 sequences are shown in Figs. 3.3 and 3.4. In all cases, the major parent was found to be most similar to one of the modern genotype 3 sequences included in the analysis, and the minor parent to resemble an ancient genotype 1 sequence most similar to DA337. To determine the timing of the recombination, we subsequently removed DA337, DA66, and a modern sequence (AB126262). Only after the removal of these three sequences was DA251, our oldest sequence, found to be the minor parent, indicating that the recombination event most likely took place between DA251 and DA337. Because of the age of the samples, the major parent clearly cannot be a modern sequence, suggesting that a genotype 3 ancestor and a genotype 1 virus that existed sometime during the evolutionary span from DA251 (6.9 kya) to DA337 (4.0 kya) recombined to form genotype 2. The identification of genotype 2 as a recombinant and the inferred recombination breakpoints closely correspond to the results of Shen *et al.,* (2016) [264] in their analysis of modern B19V sequences (Figs. 3.3 and 3.4).

Recombination algorithm	Number of sequences detected in (DA337)	Average P value (DA337)	Number of sequences detected in (DA251)	Average P value (DA251)
RDP	8	8.043×10^{-7}	7	1.037×10^{-7}
GENECONF	2	1.125×10^{-2}	4	4.386×10^{-4}
BootScan	8	5.227×10^{-13}	7	3.517×10^{-7}
MaxChi	8	3.281×10^{-10}	7	5.144×10^{-8}
Chimera	8	3.940×10^{-10}	7	1.770×10^{-9}
SiScan	8	1.349×10^{-12}	7	1.293×10^{-11}
3Seq	8	1.265×10^{-11}	7	1.032×10^{-10}

Table 3.3: Recombination analysis number of sequences and P values. There were 8 genotype 2 sequences in the recombination analysis. Columns 2 and 3 show the number of genotype 2 sequences in which the recombination was detected (maximum 8) and the average P value when the genotype 1 sample DA337 (4.0 kya) was selected as the minor parent. When genotype 1 sequences DA337, DA66, and the modern sequence AB126262 are removed from the analysis, DA251 (6.9 kya) is chosen as the minor parent, also with strong P values (columns 4 and 5). Recombination breakpoints and confidence intervals are given in the legend of Fig. 3.4. *Table contributed by Terry Jones.*

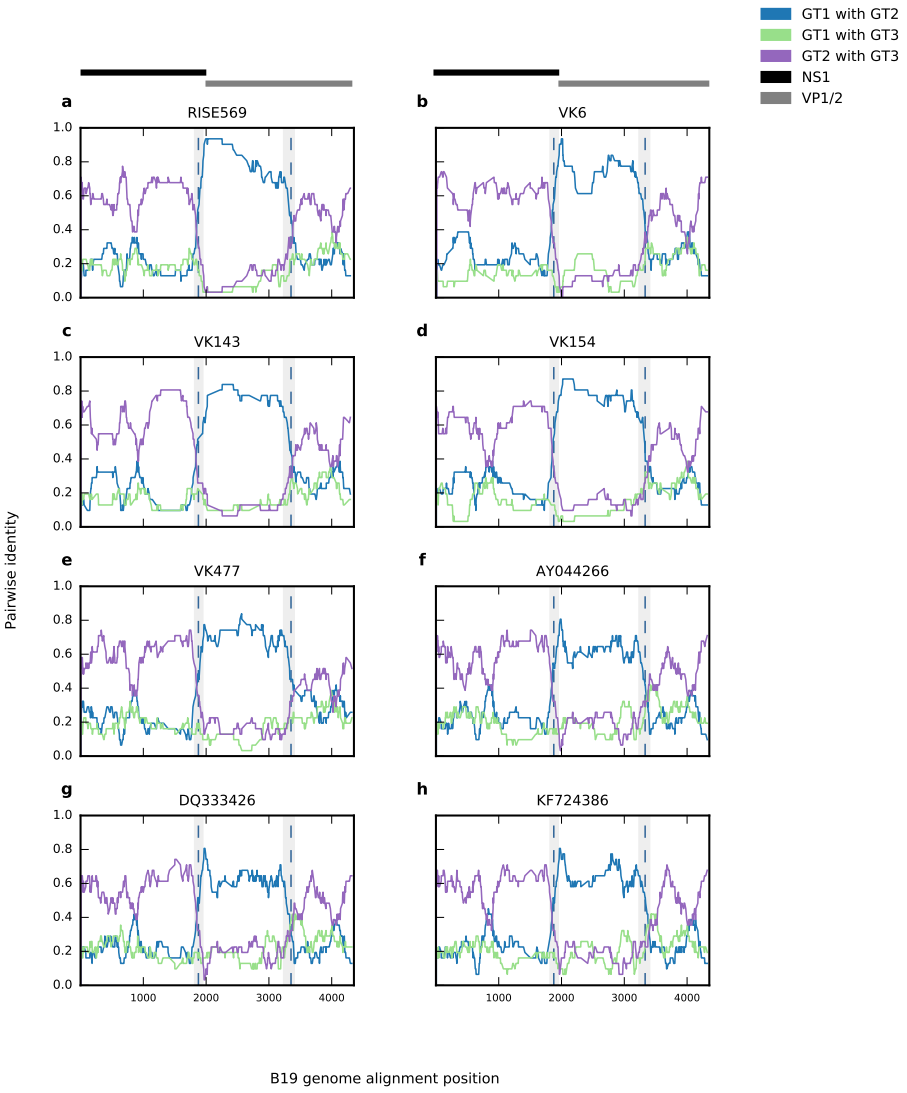


Figure 3.3: Recombination analysis using the RDP algorithm in the RDP4 package for eight genotype 2 sequences. Recombination analysis using RDP4 [214] was performed on all 10 ancient sequences and a selection of 13 modern genotype 1–3 sequences. All seven recombination detection algorithms used by RDP4 identified all genotype 2 sequences (including the 5 ancients) as potential recombinants, with strong P values (Table 4.3). In all cases, the major parent was suggested to be a genotype 3 sequence and the minor parent an ancient genotype 1 sequence. Values for the predicted breakpoints (shown as vertical dashed lines) and 99% confidence intervals (shaded in grey) are given in Fig. 3.4. The horizontal bars at the top of the figures on the first line indicate the positions of the NS1 and VP1/2 genes in the genome. *Figure contributed by Terry Jones.*

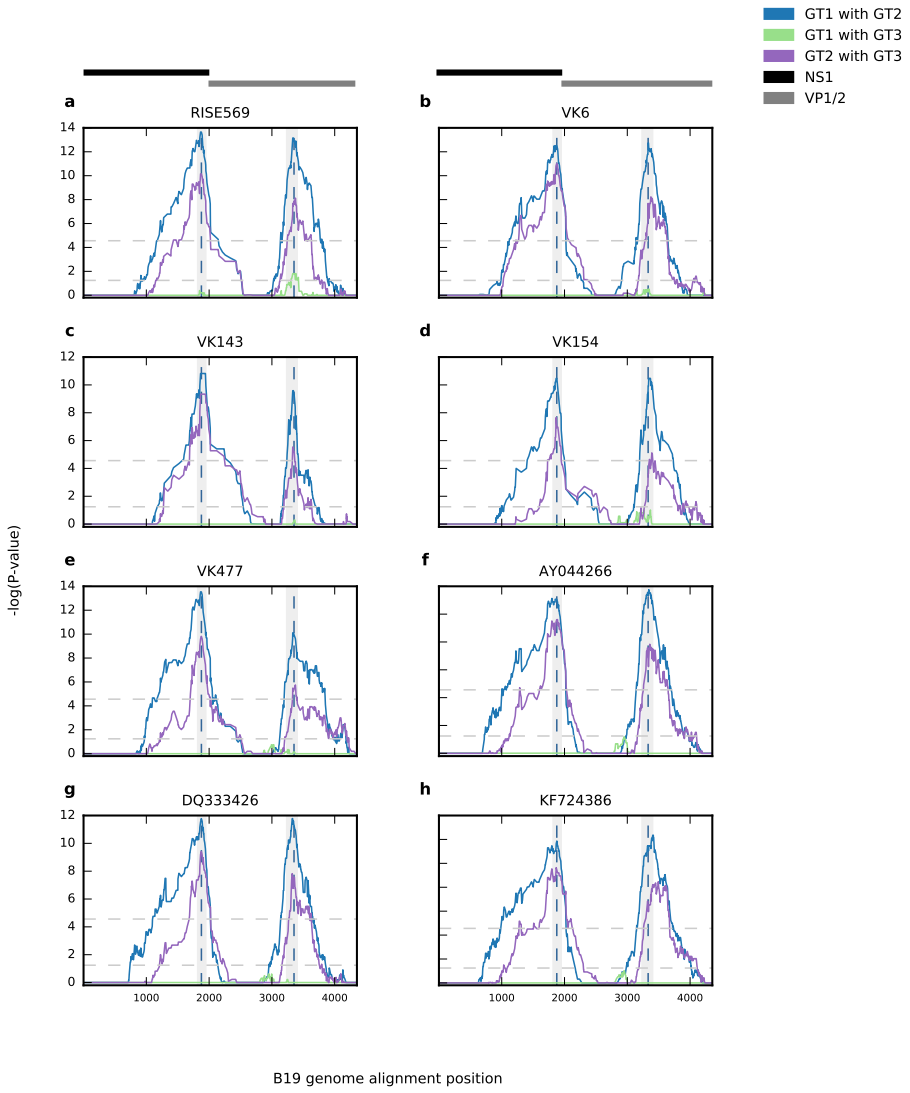


Figure 3.4: Recombination analysis using the MaxChi algorithm in the RDP4 package for eight genotype 2 sequences. The start and end breakpoints were predicted as AY044266 (1877–3352), DQ333426 (1877–3324), KF724386 (1878–3388), RISE569 (1876–3332), VK6 (1838–3332), VK143 (1902–3332), VK154 (1877–3352), VK477 (1877–3332). In all cases, the 99% confidence interval for the start breakpoint was 1822–1943. For all cases except VK477 (2885–3034), the 99% confidence interval for the end breakpoint was 3239–3399. In all cases, the 99% confidence interval (shaded in grey) of the inferred recombination breakpoints (vertical dashed lines) includes the start and end values (1889 and 3280) found by (Shen *et al.,* (2016) [264]) who also identified genotype 2 as a genotype 1/3 recombinant. The horizontal bars at the top of the figures on the first line indicate the positions of the NS1 and VP1/2 genes in the genome *Figure contributed by Terry Jones.*

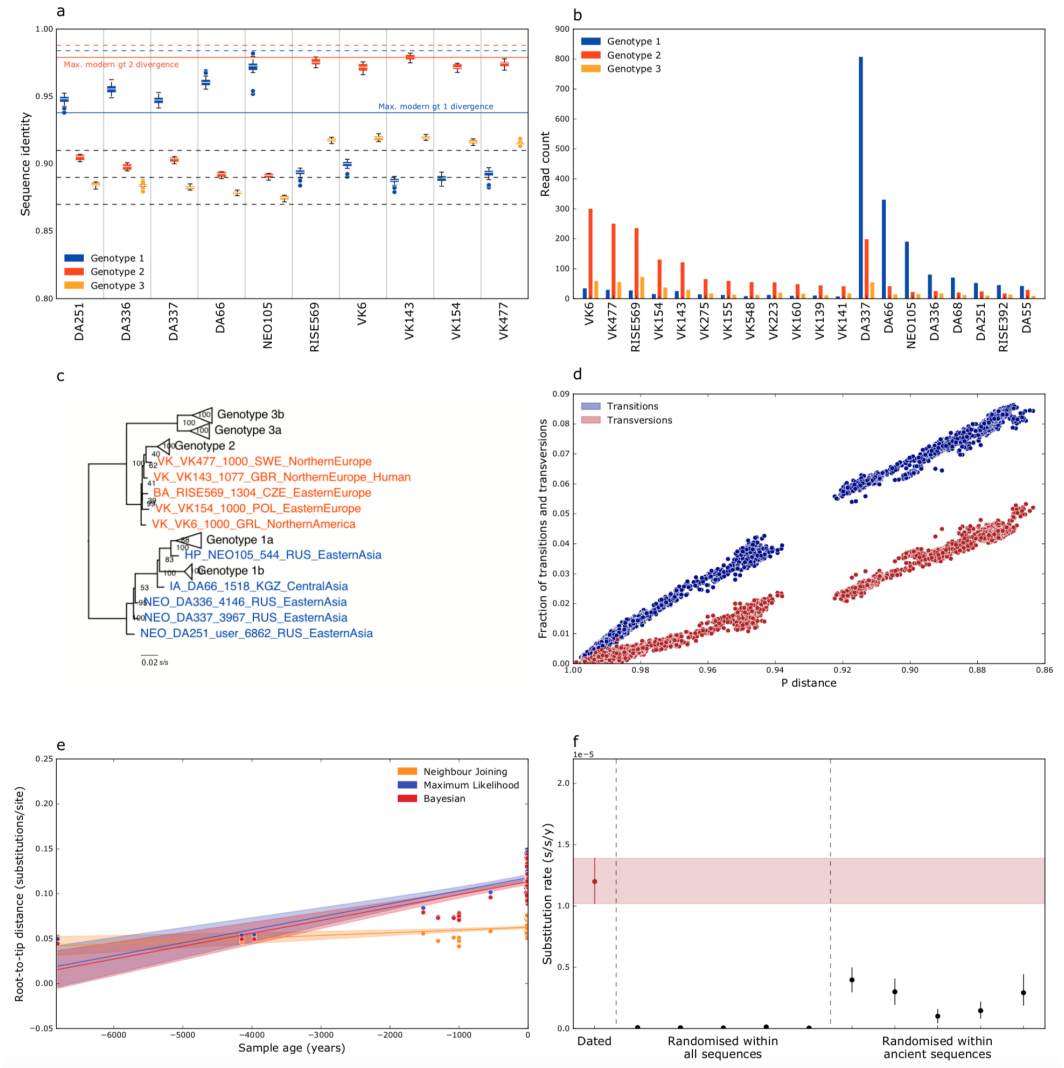


Figure 3.5: Phylogenetic analyses and sequence similarity. a) Pairwise sequence identity between ancient and modern sequences by genotype. Dashed and continuous blue and red lines denote the mean and maximum sequence divergence within a modern genotype, respectively. Black dashed lines show the mean sequence divergence between genotypes (genotype 1 vs genotype 2: 0.89; genotype 1 vs genotype 3: 0.87; genotype 2 vs genotype 3: 0.91). **b) Number of reads matching against each genotype. c) Maximum likelihood (ML) tree.** The ML tree was inferred using PhyML, using the TN93 substitution model and optimizing for topology, branch lengths, and rates, with 100 bootstrap replicates (see Methods). The x-axis denotes substitutions per site (s/s). Ancient genotype 1 sequences are shown in blue and ancient genotype 2 sequences in red. Taxon names indicate: dataset name, sample name, age, country abbreviation of sequence origin and region of sequence origin. **d) Pairwise transition and transversion frequencies plotted against P distance (sequence identity).** If saturation occurs, the transversion frequencies should be higher than the transition frequencies at large P distances. This is not the case here, and no signal of saturation can be detected.

Figure 3.5: cont. e) Regression of root to tip distances and sample age. Branch lengths were inferred using neighbor joining, maximum likelihood, and Bayesian methods. Root-to-tip distances were extracted using TempEst [225], selecting 'best-fitting-root' and the regression analysis was performed in SciPy [226]. Shaded areas represent the 95% confidence. The slopes are $3.0x10^{-6}$, $1,4x10^{-5}$, and $1.4x10^{-5}$ and the correlation coefficients are 0.38 (R^2=0.15), 0.66 (R^2=0.44), and 0.7 (R^2=0.48) for NJ, ML, and Bayesian methods, respectively. **f) Date randomization tests.** The rate and 95% HPD interval inferred using correctly dated sequences is shown in red. Two types of randomization were performed in replicates of 5, either randomizing tip dates of all sequences, or randomizing tip dates of just the ancient sequences. None of the 95% HPD intervals of the randomized runs overlap with the 95% HPD intervals of the correctly dated run, suggesting a good temporal signal in the data.

3.4.4 Phylogenetic analysis

A phylogenetic tree inferred using maximum likelihood methods confirmed that the ancient sequences fall within the diversity of known B19V sequences (Fig. 3.5c). Five ancient sequences (DA66, DA251, DA336, DA337, NEO105) fall within or basal to genotype 1, and five (RISE569, VK6, VK143, VK154, VK477) fall basal to genotype 2, consistent with the sequence similarity result presented earlier. Substitution saturation is known to affect the phylogenetic signal, and hence the inference of phylogenetic trees. However, by plotting transition and transversion frequencies against genetic distance (Fig. 3.5d) or by testing for substitution saturation using DAMBE (Table 3.4) [262], we did not find evidence for saturation in our sequences.

numOTU	Iss	Iss.cSym	T	DF	P	Iss.cAsym	T	DF	P
4	0.211	0.85	90.675	3042	0	0.84	89.314	3042	0
8	0.192	0.844	98.212	3042	0	0.764	86.067	3042	0
16	0.199	0.83	98.869	3042	0	0.676	74.731	3042	0
32	0.189	0.81	102.134	3042	0	0.561	61.139	3042	0

Table 3.4: Results from testing for substitution saturation in DAMBE. Tests for substitution saturation were performed in DAMBE, according to Xia *et al.,* (2017) [261]. The proportion of invariant sites was set to 0.3. If the Iss is smaller than Iss.cSym and Iss.cAsym, there is no evidence for saturation, as is the case here.

To infer dated coalescent trees, sufficient temporal signal must be present in the data. A regression of root-to-tip distances from trees inferred using neighbor joining, maximum likelihood, and Bayesian methods all showed a temporal signal in the data (Fig. 3.5e). Date randomization tests [228, 265] performed in BEAST2 [150] also support

the presence of a temporal signal (Fig. 3.5f). Dated coalescent trees were consequently inferred using BEAST2 (Fig. 3.6) [150]. In the following text, 95% highest posterior density (HPD) intervals are given in parentheses. We inferred a substitution rate of 1.22×10^{-5} (1.04×10^{-5}–1.40×10^{-5}) s/s/y and a root age of 10.1 (9.0–11.3) kya under a strict clock and a coalescent Bayesian skyline population prior (Table 3.5) and a substitution rate of 1.67×10^{-5} (1.24×10^{-5}–2.14×10^{-5}) s/s/y and a root age of 8.4 (7.0–11.1) kya under a relaxed log-normal clock and a coalescent Bayesian skyline population prior, with identical topologies. The time to the most recent common ancestor of genotypes 1, 2, and 3 are 7.1 (6.9–7.3), 1.9 (1.7–2.1), and 2.5 (2.1–3.0) kya, respectively, under a strict clock and 7.3 (6.9–7.9), 1.7 (1.4–2.0), and 1.5 (0.8–2.4) kya, respectively, under a relaxed log-normal clock.

Model (clock, population)		Strict, constant	Strict, exponential	Strict, Bayesian skyline
	likelihood	-25519.8	-25516.8	-25470.1
Strict, constant	-25519.8	0	2.97	49.7
Strict, exponential	-25516.8	-2.97	0	46.73
Strict, Bayesian skyline	-25470.1	-49.7	-46.73	0

Table 3.5: Model testing for different population priors. Models were compared using path sampling, as implemented in BEAST2. Log marginal likelihood values were compared using a Bayes factor test. A positive value for the Bayes factor implies support for the column model, a negative value support for the row model. A Bayes factor in the range of 3–20 implies positive support, 20–150 strong support, and >150 overwhelming support [229]. 'likelihood' refers to the log marginal likelihood estimate. Abbreviations are as follows: 'Strict': strict clock, 'Relaxed-log.': relaxed log-normal clock, 'Constant': coalescent constant population prior, 'Exponential': coalescent exponential population prior, 'Skyline': coalescent Bayesian skyline population prior.

Recombination is known to affect the accuracy of phylogenetic analyses. Hence, to better understand the effect of recombination on the dated B19V tree, we inferred trees under a strict clock and coalescent Bayesian skyline population prior for the three genotypes individually, dating the most recent common ancestor of genotypes 1, 2, and 3 to 6.9 (6.9–6.9) kya, 1.6 (1.4–1.8) kya, and 0.6 (0.1–6.1) kya, respectively. Further, we inferred trees under a strict clock and coalescent Bayesian skyline population prior for the full genome, but excluding all genotype 2 sequences (Fig. 3.7b), and separately for the region of the minor (Fig. 3.7c) and major (Fig. 3.7d) parent. The maximum root age was inferred to be 12.6 (10.4–15.2) kya when only the minor parent was considered (Fig. 3.7c); however, the 95% HPD interval of the root ages overlapped in all cases. The 95% HPD interval on the split of genotype 2 from genotype 1 in Fig. 3.7c) suggests that the recombination event that formed genotype 2 occurred between 5.0–6.8 kya.

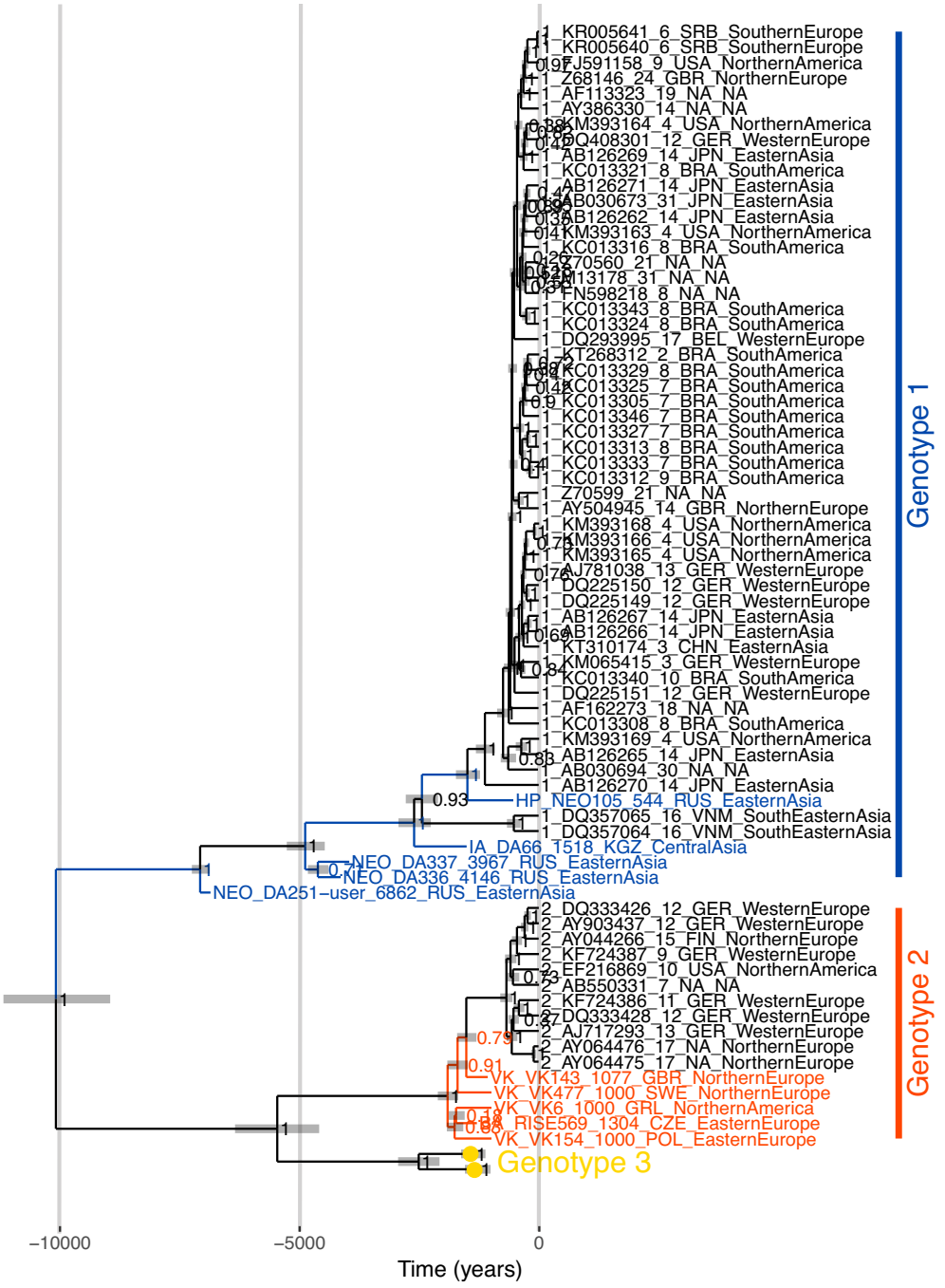


Figure 3.6: Maximum clade credibility tree inferred in BEAST2. A strict clock and coalescent Bayesian skyline population prior were used. The substitution rate is inferred as 1.22×10^{-5} s/s/y (95% HPD interval, 1.04×10^{-5}–1.40×10^{-5} s/s/y), and the root age as 10.1 kya (95% HPD interval, 9.0–11.3 kya). The x-axis denotes time into the past, in years. Taxon names: genotype/historical period, accession number/sample identifier, sample age, country abbreviation of sequence origin and region of sequence origin, as determined by the Standard country or area codes for statistical use (42). Horizontal grey bars indicate 95% HPD intervals of node ages.

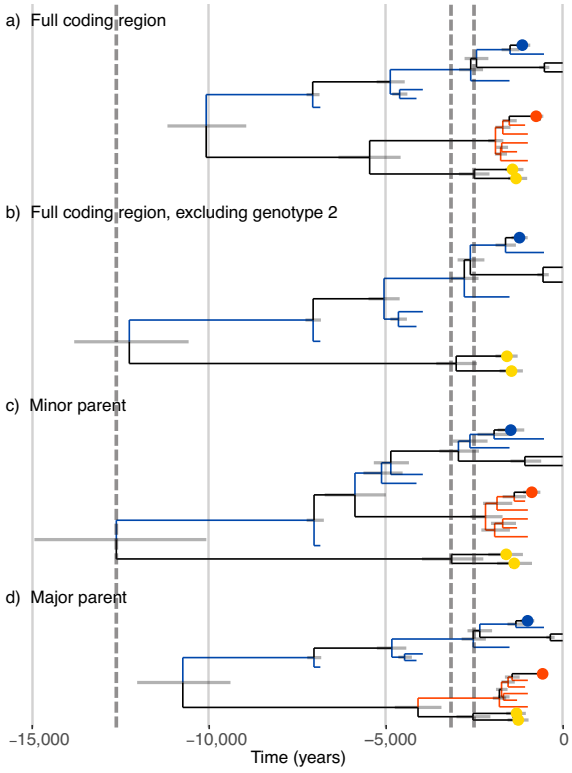


Figure 3.7: Trees inferred using BEAST2 from different regions of the B19V genome. Trees were inferred using a strict clock and coalescent Bayesian skyline population prior. Positions of ancient genotype 1 and ancient genotype 2 sequences are shown as blue or red tips, respectively. Collapsed modern genotype 1, 2, and 3 sequences are shown as blue, red, and yellow dots, respectively. 95% HPD intervals of node ages are shown by grey horizontal bars. The x-axis denotes time into the past, in years. Vertical grey dashed lines indicate, from left to right, the maximum height of the root and the maximum and minimum height of the genotype 3 clade. **a) Tree from Fig. 2 using the full coding region. b) Tree using the full coding region, excluding genotype 2 sequences. c) Tree using the minor parent** (positions 1,877–3,352 of the alignment). **d) Tree using the major parent** (positions 1–1,876 and 3,353–4,354 of the alignment).

3.5 DISCUSSION

We show that it is possible to recover and sequence single-stranded ancient viral DNA molecules from human skeletons that are up to ~6.9 thousand years old. Although Toppinen *et al.,* (2015) [126] found that human B19V DNA can persist in human bone for decades, the fact that ssDNA is able to persist for ~6.9 kya may seem surprising, given that depurination and deamination occur about four times faster in ssDNA than in double-stranded DNA [64]. However, single-stranded B19V DNA may be protected by the viral capsid, by self-annealing of ssDNA during preservation, or through the binding of ssDNA to the inside of the viral capsid, as observed in canine Parvovirus [266]. This finding also provides hope for the retrieval of RNA viruses from archaeological samples, notwithstanding the traditional view of RNA as too unstable for long-term survival [64].

By computational screening of shotgun data from 1,578 ancient individuals, we found 20 samples with reads matching 30–99.7% of the B19V coding region, eight of which had a majority of reads matching genotype 1 and 12 matching genotype 2. No sample had a majority of reads matching genotype 3. The phylogenetic positioning of the ancient genotype 1 sequences may suggest a ladder-like evolution, in which older virus strains go extinct as they are replaced by novel variants, as observed in modern genotype 1 sequences [158].

The ancient sequences allow us to revise the current timelines for B19V evolution. Specifically, we show that B19V has been associated with humans for thousands of years, given our finding of ancient sequences as old as ~6.9 kya that are still most similar to genotype 1. This rules out the suggestion that genotype 1 only arose after World War II [126,158]. However, endogenous parvovirus-derived genomes found in a large range of mammals suggests that the virus may be considerably older than 12.6 kya [165]. Furthermore, we find a substitution rate of 1.22×10^{-5} s/s/y, approximately an order of magnitude lower than rates inferred solely from modern B19V sequences [126, 157, 158]. Our rate estimate is in line with rates estimated for other single- and double-stranded DNA viruses [35]. The fact that we find lower rates when sampling over the course of a longer period, as opposed to sampling over the course of a shorter period, is consistent with the phenomenon of time dependency of substitution rates [152,162]. The most likely cause of this phenomenon are the inability of evolutionary models to account for saturation (Peter Simmonds, personal communication). Other causes, such as changes in selection pressure and virus biology, as well as transient mutations have been mentioned [161]. Since viral substitution rates decay gradually with increasing measuring period, changes in selection pressure and virus biology are

deemed unlikely [38, 152]. Inferred B19V substitution rates may decrease as older ancient sequences are discovered and incorporated into future analyses.

It has been shown previously that genotype 2 was formed by a recombination between genotypes 1 and 3 [264]. Using the position of the ancient sequences as shown in Fig. 3.7c), we were able to date this recombination event to between 5.0 and 6.8 kya. Recombination leads to the underestimation of root ages in dated coalescent trees, which is what we observed when comparing trees inferred from the full genomes (Fig. 3.7a) with trees inferred from only genotypes 1 and 3 (Fig. 3.7b) or from the minor parent (Fig. 3.7c). Interestingly, we did not find the same when inferring a tree using the major parent (Fig. 3.7d), where the root age is underestimated, similar to the full tree in Fig. 3.7a). However, in all cases, the 95% HPD intervals of the root ages overlap.

On the basis of our data, it is possible to hypothesize a geographical separation between the ancient genotype 1 sequences, found in Central Asia, and ancient genotype 2 sequences, found in Europe and Greenland. However, this may simply be a result of sampling bias due to the small number of positive samples. Although no ancient genotype 3 sequences have been found, there appears to be no feature inherent to the biology of genotype 3 that inhibits long-term preservation of its DNA [258], as the genotype has been found in remains of World War II soldiers by Toppinen *et al.,* (2015) [126].

This small number of ancient samples of B19V has allowed us to revise previous best estimates of rate of evolution and the timespan of the association of B19V with humans, both by an order of magnitude. Similar gains in our understanding of the evolutionary processes of other pathogens, and for improving phylogenetic inference methods when no ancient sequences exist, surely await.

CHAPTER 4: DIVERSE VARIOLA VIRUS LINEAGES CIRCULATED IN NORTHERN EUROPE DURING THE VIKING AGE

PREFACE

A version of this chapter is currently in review at Science as (* denotes equal contribution):

Barbara Mühlemann*, Martin Sikora*, Lasse Vinner*, Ashot Margaryan, Helene Wilhelmson, Constanza de la Fuente Castro, Morten E. Allentoft, Peter de Barros Damgaard, Anders Johannes Hansen, Sofie Holtsmark Nielsen, Lisa Mariann Strand, Jan Bill, Alexandra Buzhilova, Tamara Pushkina, Ceri Falys, Valeri Khartanovich, Vyacheslav Moiseyev, Marie Louise Schjellerup Jørkov, Østergaard Sørensen, Hannes Schroeder, Gerd Sutter, Geoffrey L. Smith, Christian Drosten, Ron A. M. Fouchier, Derek J. Smith, Terry C. Jones, Eske Willerslev. *Diverse variola virus lineages circulated in northern Europe during the Viking Age*.

It has been modified to fit the style of a dissertation.

The data was screened and authenticated in parallel by Terry Jones and myself, as well as by Martin Sikora. I assembled the consensus sequences, did the phylogenetic analyses (except the placement of the low-coverage samples), gene inactivation analysis (with expert input from Geoffrey Smith), figures (except the placement of the low-coverage samples) and tables (except the recombination table), and wrote manuscript in collaboration with Terry Jones, Derek Smith, Geoffrey Smith, Gerd Sutter, Ron Fouchier, and Christian Drosten, and with input from all co-authors. The recombination analysis was done by Terry Jones, and the trees with the low-coverage samples were made by Martin Sikora. Capture of a subset of samples was performed by Lasse Vinner. Original sequencing was performed by Morten Allentoft, Ashot Margaryan, Peter de Barros Damgaard, and Constanza de la Fuente Castro. In addition to this description, I have noted in the legends of figures and tables if they were contributed by others.

4.1 ABSTRACT

We present thirteen cases of smallpox from humans living in northern Europe, including eleven from the Viking Age (~800–1100 Common Era). Viral sequences isolated from these individuals reveal a now-extinct sister clade to the variola viruses in circulation prior to the eradication of smallpox. We date the most recent common ancestor of all now known variola viruses to ~1,600 years ago. A diverse pattern of gene inactivation shows that the Viking age viruses were following multiple parallel evolutionary paths. Sixteen genes disrupted in modern variola viruses are intact in the Viking age viruses, nine of which are immune modulators that affect virulence in the related vaccinia virus. Identical gene-inactivating mutations in five other genes indicate a single zoonotic introduction of variola virus into humans. These findings reveal a richer and more diverse genetic history of variola virus than previously suspected.

4.2 INTRODUCTION

Variola virus, the causative agent of smallpox, is estimated to have caused 300–500 million deaths in the 20th century alone [10]. Despite the eradication of smallpox, there are ongoing concerns regarding the re-emergence of a smallpox-like disease, either due to accidental or deliberate re-introduction of variola virus, adaptation of monkeypox virus to man, or via zoonosis or genetic engineering of another orthopoxvirus [267–269]. Thus, a better understanding of the evolutionary history of variola virus is of substantial interest.

The orthopoxviruses are a genus of the *Poxviridae* and have large, linear, double-stranded DNA genomes [270]. They differ in the range of mammalian host species they infect and the severity of the disease they cause: Variola virus (VARV) and camelpox virus (CMLV) have a narrow host range and can be highly virulent, while taterapox virus (TATV), infecting gerbils, does not cause significant morbidity [10, 271]. Cowpox virus (CPXV), monkeypox virus (MPXV), and vaccinia virus (VACV) infect several host species, with varying severity of disease. VACV, whose origin and reservoir host is unknown [272], generally has low virulence in humans, elicits an immune response that is protective against VARV, and was used as the vaccine during the smallpox eradication campaign [10]. VACV infection of mice is a useful model for studying orthopoxvirus gene function in vivo. Orthopoxvirus genomes are typically between ~186,000 and ~220,000 nucleotides (nt) long and contain between 179 and 212 genes [10]. Conserved genes, necessary for virus transcription and replication, are located in the central ~100,000 nt region of the genome, and are commonly used in phylogenetic inference. In vitro and in vivo studies show that orthopoxvirus genomes also contain many genes important for modulating host innate immunity and determining host range, but whose deletion does not prevent virus replication [7, 273]. Those are located near the genome termini, and are variable between different orthopoxvirus species. The factors governing host range and virulence of orthopoxviruses are complex and not fully understood. For example, some genes that promote VACV virulence are inactivated in the more virulent VARV [274–276], while in other cases the loss or inactivation of host immune system-modulating genes in VACV can result in increased virulence [7].

The timeline of the emergence of smallpox in humans is unclear. Analysis of sequences from a 17th century Lithuanian mummy (VARV-VD21) [116], two Czech museum specimens from the 19th and 20th century [115], and VARV from the 20th century sampled during the eradication campaign, dates their most recent common ancestor (MRCA) to the 17th or 18th centuries [115, 116, 277, 278]. However, written

records of possible smallpox infections date back to at least 3000 years ago (ya), and mummified remains suggestive of smallpox date to 3570 ya [279, 280]. Due to the absence of sequences older than ~360 years, a large gap remains in our knowledge of the evolution of VARV.

Here we present three high-coverage and 10 low-coverage VARV sequences from northern Europe. Eleven of the samples (including the three with high-coverage) are dated between ~550–1100 Current Era (CE), and two between 1800–1900 CE.

4.3 MATERIAL AND METHODS

4.3.1 Datasets

Ancient sequences:
aVARV-VK382, aVARV-VK388, aVARV-VK470.

Dataset 1 (Human tree):
51 VARV sequences plus the ancient sequences listed above.

L22579, DQ441420, DQ441421, DQ437581, DQ437588, DQ441417, DQ441418, DQ441435, DQ441436, DQ441427, DQ441446, DQ441445, DQ441444, DQ437584, DQ441428, DQ441429, DQ441430, DQ441431, DQ441432, DQ441442, DQ437591, DQ441424, DQ441425, DQ441438, DQ441439, DQ437590, DQ441440, DQ441441, DQ437582, DQ441433, DQ437585, DQ437586, DQ441448, DQ437580, DQ437587, DQ437592, DQ437589, DQ441423, DQ441443, DQ437583, Y16780, DQ441419, DQ441447, DQ441416, DQ441426, DQ441437, DQ441434, NC_001611, V563: LT706528, V1588: LT706529, VARV-VD21: BK010317

Dataset 2 (full Maximun likelihood tree):
Dataset 1 plus the following sequences (abbreviations: TATV: Taterapox virus, CMLV: Camelpox virus, ECTV: Ectromelia virus, CPXV-Gri: Cowpoxvirus Gri, CPXV-Ger: Cowpoxvirus Germany, CPXV-BR: Cowpoxvirus Brighton Red, VACV: Vaccinia virus, VACV-Ank: Vaccinia virus Ankara, VACV-COP: Vaccinia virus Copenhagen, MPXV-Zai: Monkeypox virus Zaire, HPXV: Horsepox virus):

TATV: DQ437594 (NC_008291), CMLV M-96: AF438165 (NC_003391) , CMLV-CMS: AY009089, ECTV: NC_004105, ECTV: KY554976, ECTV: JQ410350, ECTV: KJ563295, CPXV: KC813508, CPXV-Gri: X94355, VACV-Ank: AM501482, CPXV: HQ420893, VACV: AY313848, VACV: KF179385, CPXV: DQ437593, VACV: KJ12-5439, CPXV: KC813493, VACV: JX489138, CPXV: HQ420895, CPXV: KC813509, CPXV: HQ420897, CPXV: KC813492, CPXV: HQ420894, CPXV: KC813511, CPXV: NC_003663, CPXV: HQ420899, CPXV: HQ420896, MPXV-ZAI: NC_003310, CMLV: KP768318, HPXV: DQ792504, MPXV: DQ011156, Akhmeta: MH607143, VACV: NC_006998, VACV-COP: M35027

Dataset 3 (References for analysis of gene-inactivating mutations):
17 sequences as described in Hendrickson *et al.*, (2010) [270].

CMLV: AF438165 (NC_003391), CPXV-BR: AF482758 (NC_003663), CPXV-Ger: DQ437593, CPXV-Gri: X94355, ECTV: AF012825 (NC_004105), MPXV-WR: AY60-

3973, MPXV-ZAI: AF380138 (NC_003310), TATV: DQ437594 (NC_008291), HSPV: DQ792504, RPXV: AY484669, VACV-MVA: U94848, VACV-COP: M35027, VACV-WR: AY243312 (NC_006998), VARV-BRZ: DQ441419, VARV-KUW: DQ441433, VARV-SLN: DQ441437, VARV-SAF: DQ441436

4.3.2 Sample preparation, screening, authenticity

Thirty-two samples were selected for enrichment of viral target sequences in library by capture, based on the availability of additional sampling material and the presence of reads matching orthopoxviruses using Kraken [281] and DIAMOND [207] (version 0.9.23) algorithms. A total of 1653 humans living in Eurasia and the Americas between ~31,630 and ~150 ya were screened. Eleven of the 32 individuals had already been enriched for viral target sequences in library capture in a previous study [79]. From the remaining samples, uniquely dual-indexed sequencing libraries (2–6 libraries/sample) were prepared from 1–3 ancient DNA (aDNA) extracts/sample. Libraries were pooled into capture reactions containing 1183–1667 ng/rxn. Hybridisation and sequencing were performed as described previously [79].

In order to confirm the presence of authentic orthopox reads, reads from the capture were mapped against a set of nine reference genomes distributed across the orthopox phylogeny using Bowtie2 [282] (version 2.3.2). For each reference genome, both end-to-end (`--end-to-end`) and local (`--local`) alignment were performed, using a modification of the `'very sensitive'` preset allowing for an additional mismatch in the seed alignment (`-N 1`). Following this approach, a sample was considered a likely positive for infection with an ancient VARV according to the following criteria:

- At least 30 unique reads with MQ30 mapping against the VARV genome.
- Mapped reads randomly distributed across the entire VARV genome *(Contributed by Martin Sikora, not shown)*.
- Lowest read edit distances against either CMLV, TATV, or VARV *(Contributed by Martin Sikora, not shown)*.

A sample was considered a confirmed positive if, in addition to the above, the reads showed clear ancient DNA damage patterns as revealed by mapDamage [209]. Furthermore, in all ancient samples, 100% to 97.4% of all reads identified as described above, had orthopoxvirus as their best match when using BLASTn [25] (version 2.8.1) to map against the entire NCBI nt database downloaded on January 4, 2019.

The fact that 10 of the 11 Viking age ancient samples group phylogenetically as a novel clade in sister relationship to modern VARV further supports authenticity.

To assemble the reads used for the generation of consensus sequences and the investigation of gene-inactivating mutations, reads from the capture were mapped against orthopoxvirus reference genomes using DIAMOND (version 0.9.23), and BWA [206] (version 0.7.15-r1140) (mem and aln (with and without the `-l` option) algorithms). Reads from all mapping algorithms were combined and de-duplicated for further use.

4.3.3 Generation of consensus sequences

Orthopoxviruses differ in their gene content [270, 274]. As we do not know which genes are present in the ancient viruses, assembling a full genome by mapping to a modern reference sequence is impossible. We therefore constructed ancient consensus sequences only for the conserved central section of the genome, from gene VACV-COP *F4L* to gene VACV-COP *A24R*. Reads were aligned to a TATV reference sequence (accession: NC_008291) using Geneious [211] (version 9), medium sensitivity, fast parameters. In the resulting alignments, we manually trimmed damaged ends, and set C/T and G/A mutations at positions with less than five reads coverage to 'N'.

4.3.4 Recombination analyses

The recombination analyses and the associated table and text in this section were contributed by Terry Jones. Seven algorithms, as implemented in the RDP4 [214] (version 4.9.5) package, were used to search for evidence of recombination in a selection of 10 orthopoxvirus sequences. The algorithms were: 3Seq [220], BootScan [216], Chimaera [218], GENECONV [215], MaxChi [217], RDP [263], and SiScan [219]. The sequences (with accession numbers in parentheses) were: Akhmeta (MH607143), CMLV (AY00-9089), CPXV (KC813492), CPXV (KC813508), TATV (NC_008291), VARV-VD21 (BK010317), VARV (NC_001611), aVARV-VK382, aVARV-VK388, and aVARV-VK470. A region of 55 nucleotides of aVARV-VK470 was identified as being of possible recombinant origin, but by only one RDP4 algorithm (GENECONV), with aVARV-VK388 as the major parent with the minor contribution from an unknown parent. The RDP4 program issued the caveat that the proposed recombination signal may be attributable to a process other than recombination. Similarly, a region of 261 nucleotides in aVARV-VK382 was flagged as a possible recom-

binant by three algorithms (3Seq, BootScan, and GENECONV), again with aVARV-388 as the major parent and a contribution from an unknown parent, and with the same caveat. No recombination signal was detected with aVARV-aVARV-VK388 as the recipient sequence.

4.3.5 Phylogenetic analyses

4.3.5.1 Maximum likelihood trees

Separate trees were made from Datasets 1 and 2. The sequences were aligned using MAFFT [221] (version 7). Maximum likelihood (ML) trees were generated using RAxML-ng [283] (version 0.7.0 BETA) with a K81uf+G+I (selected using bModelTest [227]) substitution model, and an ML estimate of tree topology, branch lengths, substitution rates, and nucleotide frequencies and support estimated using 1000 bootstrap replicates.

4.3.5.2 Low-coverage samples trees

Analyses described in this section were performed by Martin Sikora. The EPA-ng algorithm [284] (version 0.3.4) was used to infer the positions of consensus sequences from the low-coverage samples (VK108, VK138, VK168, VK255, VK281, VK359, VK515, VK533, FIN1, and KHA1) on the ML tree made from Dataset 2, including the sequences from the three high-coverage samples. For all samples, consensus sequences were obtained by extracting the majority allele from BAM files mapped against a VARV reference genome (accession number NC_001611.1) using the local alignment approach, to reduce the impact of genotyping errors due to post-mortem DNA damage (Fig. 4.2). The resulting consensus sequence was aligned to the multi-sequence alignment of Dataset 2 from which the original ML tree was inferred (see 'Maximum likelihood trees', above) using MAFFT (version 7.407) with the `--add` and `--keeplength` options. EPA-ng was run using the same model parameters as used for the original ML tree (K81uf+G+I). Placements of low-coverage samples were further analyzed using gappa [285] (version 0.1.0), and the uncertainty in placements for each sample was visualized using likelihood weight ratios (LWRs) on the reference tree using the ggtree [286] (version 1.14.6) package in R [287] (version 3.5.2).

4.3.5.3 Dated coalescent trees using BEAST2

We performed a linear regression of root-to-tip dates against sampling dates. Root-to-tip distances were extracted using TempEst [225] (version 1.5) from the ML tree inferred from Dataset 1 and the sequences from the three high-coverage samples in Fig. 4.3 (Dataset 1). The regression analysis was performed in SciPy [226]. Dated coalescent trees were inferred using BEAST2 [150] (version 2.4.7). Forty-eight modern VARV sequences were used, as well as two published ancient sequences from Czech Republic, and VARV-VD21 (Dataset 1). Using bModelTest [227] (version 1.0.4), we selected a TPM1 substitution model with unequal base frequencies, invariant sites, and gamma distributed rate heterogeneity among sites. The prior on the clock rate was constrained using a uniform($1x10^{-9}$–$1x10^{-3}$ substitutions / site / year (s/s/y)) prior. Proper priors were used throughout. Trees were inferred using strict and relaxed log-normal molecular clocks, as well as constant and exponential coalescent and Bayesian skyline population priors. The Markov chain Monte Carlo analysis was run for 100M generations, sampling every 2000 generations. An effective sample size >200 was reached for all relevant parameters, and convergence and mixing was assessed using Tracer [230] (version 1.6.0). Final tree files were sub-sampled to contain 10,000 trees and Maximum Clade Credibility trees were summarised using TreeAnnotator (version 2.2.4 prerelease). Path sampling, as implemented in BEAST2, was used to select the best fitting molecular clock and population prior. Per path sampling run, 50 steps with a chain length of 1,000,000 generations were run. Likelihood values were compared using a Bayes factor test [229].

4.3.5.4 Investigating gene-inactivating mutations

In order to assess gene content in absence of a reference genome, we investigated the presence of gene-inactivating mutations in the ancient samples by comparison of alignments made to 17 orthopoxviruses. Information about the location of homologous genes in 17 orthopoxvirus reference genomes (Dataset 3) is presented in Supplementary Table 1 of Hendrickson *et al.* (2010) [270]. This was used to acquire sequences of homologous genes of those 17 reference sequences. Excluding cases where a gene is absent in a reference sequence, this leads to 3422 sequences for 219 genes. Supplementary Table 1 in Hendrickson *et al.*, (2010) [270] was also used to acquire information on the presence, absence, fragmentation, and truncation of genes in the 17 reference sequences. Reads for each ancient sample were matched against each reference gene sequence using BLASTn [25], and consensus sequences for each gene were generated using BWA [206] (version 0.7.15-r1142-dirty) (aln al-

gorithm) and samtools [288] (version 1.9). Only consensus sequences which covered more than 50% of the reference gene sequence and whose median matching read bit score was within 2 of the highest median bit score (so as to avoid excluding matches with almost-identical bit scores) among all available reference genes were considered further. Ancient consensus sequences and read alignments for each gene were manually inspected for the presence of stop codons, and insertion and deletions that indicated the inactivation of the gene in Geneious. Gene-inactivating mutations were classified as either certain or uncertain. The criteria for a certain gene-inactivating mutation were: 1) There is a mutation that leads to a stop codon (insertion / deletion / point mutation) which is covered by more than four reads. 2) The consensus sequence is against a reference that is truncated or fragmented. The criterium for an uncertain gene-inactivating mutation was: 1) There is a mutation that leads to a stop codon which is covered by less than or equal to four reads. Seeing as the filtering often returned more than one possible consensus sequence per gene and per sample, the information about gene-inactivating mutations was aggregated across multiple consensus sequences. Thus, overall, a gene-inactivating mutation in the ancient sequence was deemed to be present if a) there was a gene-inactivating mutation in all the consensus sequences for the best-matching reference genes, or b) there was a gene-inactivating mutation in some and an uncertain gene-inactivating mutation in all other consensuses. Gene-inactivating mutations were deemed to be uncertain if all consensuses had an uncertain inactivating mutation, or if one or some consensuses had an inactivating mutation but the other(s) did not. If only some consensuses had an uncertain gene-inactivating mutation, and there was no gene-inactivating mutation in others, the gene was deemed to be present. Figures 4.7a-d show all genes considered and the ancient gene consensus sequences that were considered further and whether or not they contained a stop codon. Figure 4.8 shows the inactivation of genes over time in mVARV, CMLV, TATV, aVARV and the internal phylogenetic tree nodes that connect them. The number of gene-inactivating mutations at each tip and internal node is plotted against time, with the ages of the internal nodes taken from the dated coalescent tree inferred using a log-normal relaxed clock and a Bayesian skyline population prior (Fig. 4.6). A gene-inactivating mutation is deemed to be present in an internal node if there is a gene-inactivating mutation in the same location in both children. If one or two uncertain gene-inactivating mutations are present in a child, the gene-inactivating mutation in the parent internal node is also shown as uncertain. Figure 4.9 provides additional information, allowing the direct comparison of the presence and absence of inactivating mutations on a per-gene basis.

4.4 RESULTS AND DISCUSSION

To investigate the earlier evolutionary history of VARV, we screened shotgun high-throughput sequencing data from skeletal and dental remains of 1653 humans living in Eurasia and the Americas between ~31,630 and ~150 ya. 555 of those were from northern Europe and western Russia and were context-dated to the Viking Age. Based on matches against orthopoxvirus reference sequences and the availability of additional sampling material, 32 were selected for enrichment of viral target sequences in library by capture, as described [79]. Thirteen individuals showed clear evidence of VARV infection (Table 4.1, Fig. 4.2). Of these 13 samples, 11 were Viking age individuals from northern Europe and western Russia, and two were from the 19th century from western Russia (Fig. 4.9a).

In three (hereafter 'high-coverage') samples (VK382, VK388, and VK470), there were sufficient reads to produce clear DNA damage patterns characteristic of ancient DNA, with elevated miscoding lesions towards the termini [57], and sufficient read coverage (>1.6x) to construct consensus sequences covering 86.8% to 99.6% of the conserved central region of the orthopoxvirus genome. The remaining 10 (hereafter 'low-coverage') samples consist of four samples (VK168, VK281, VK533, and FIN1) with clear DNA damage patterns but low coverage (<0.74x), and six likely-positive samples (VK108, VK138, VK255, VK359, VK515, and KHA1) whose read counts were too low to produce clear DNA damage patterns (Table 4.1, Fig. 4.2). The consensus sequences for the three high-coverage samples were collectively most similar to each other, and then to TATV (Table 4.2).

Before inferring phylogenetic trees based on the conserved central region of the genome, we tested whether modern VARVs show evidence of genomic segments that descend from a virus population containing the Viking age VARVs and whether the high-coverage Viking age VARV sequences were themselves recombinants. We examined the Viking age sequences and six representative sequences from related modern orthopoxvirus species for evidence of recombination using seven algorithms from the RDP4 toolset [214]. Although two short regions (261 and 908 nts of the ~100,000 nt central region of the VARV genome) of the high-coverage sequences were flagged as possibly due to recombination with other unknown ancient sequences, the evidence of any recombination was highly uncertain. None of the modern sequences in the selection were produced by recombination with any other modern or ancient sequence from the selection (Table 4.3).

Sample	Date (CE)	Site	Sex	Individual age (years)	Sample type	Clear DNA damage	Coverage category	Read-count	Coverage of genome	Coverage depth
VK388	550–750	Nordland, NOR	M	12–17	Tooth	Yes	High	20,585	0.990	8.13x
VK470	900–1100	Gnezdovo, RUS	F	Unknown	Tooth	Yes	High	19,069	0.971	7.92x
VK382	700–800	Öland, SWE	M	Over 15	Tooth	Yes	High	5,980	0.842	2.37x
VK281	885–990	Zealand, DEN	M	30–35	Tooth	Yes	Low	1,910	0.499	0.81x
VK168	880–1000	Oxford, GBR	M	16–25	Petrous	Yes	Low	1,965	0.428	0.7x
VK533	800–1050	Öland, SWE	*	50–60	Tooth	Yes	Low	611	0.207	0.25x
FIN1	1800–1900	Varzino, RUS	F	40–45	Tooth	Yes	Low	95	0.023	0.03x
VK515	950–1025	Nordland, NOR	M	18–22	Tooth	No	Low	124	0.03	0.04x
VK359	700–800	Öland, SWE	M	Over 15	Tooth	No	Low	62	0.013	0.01x
KHA1	1800–1900	Yamalo-Nenetskiy, RUS	N/D	3–4	Petrous	No	Low	38	0.012	0.01x
VK108	800–1000	Malmo, SWE	F	Unknown	Tooth	No	Low	55	0.01	0.01x
VK138	approx. 1000	Fyn, DEN	M	25–35	Tooth	No	Low	41	0.01	0.01x
VK255	950–1000	Gnezdovo, RUS	F	Unknown	Tooth	No	Low	30	0.01	0.01x

Table 4.1: Overview of samples with reads matching orthopoxviruses. From left to right, the columns refer to the sample name, approximate sample date as estimated from the archaeological context, the site where the sample was found (DEN: Denmark, GBR: Great Britain, NOR: Norway, RUS: Russia, SWE: Sweden), the sex of the individual, the age of the individual at the time of death, the type of tissue that was sequenced, whether a clear DNA damage pattern is evident in the sample sequencing reads (Fig. S1), and whether genome coverage is high (>1.6x) or low (<0.74x) (Fig. S2). N/D: not determined. Eleven of the samples date from the Viking Age (~800–1100 CE, with sample names starting with 'VK') or immediately prior to it and are from Scandinavian countries, western Russia, or the UK. Rows are ordered by decreasing coverage depth. *VK533 was identified as female by osteology, but is genetically male.

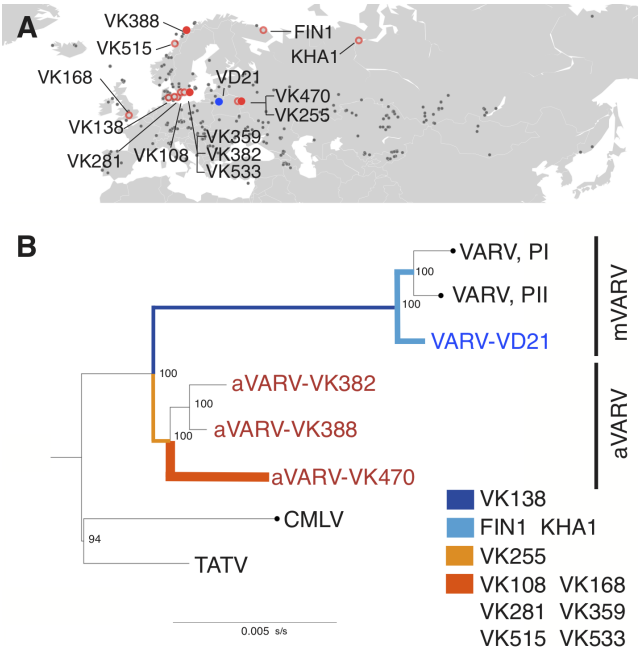


Figure 4.1: Geographical location and phylogenetic placement of the ancient samples. a) Map showing the geographical location of the samples. The high-coverage sample locations are indicated with solid red circles and low-coverage sample locations with open red circles. The location of published sample VARV-VD21 is shown in blue. Grey dots indicate locations of screened samples in which VARV was not detected (samples from South-East Asia and the Americas are omitted for clarity). Eight of the Viking age samples were from Denmark, Norway, and Sweden (Table 4.1). **b) Maximum likelihood tree showing the placement of the low-coverage samples.** Low-coverage samples were placed onto the tree using EPA-ng. Clades that do not have any low-coverage samples placed on them are collapsed and indicated by black circles. The full tree is shown in Fig. 4.3. The coloured branches indicate the branch with the highest likelihood weight ratio for each low-coverage sample. Branch thickness indicates the number of low-coverage samples placed on the branch by EPA-ng. The scale bar indicates substitutions per site (s/s).

	Closest CMLV	Closest TATV	Closest modern VARV	VARV-VD21	aVARV-VK382	aVARV-VK388
VARV-VD21	0.986	0.988	0.998	1		
aVARV-VK382	0.991	0.994	0.99	0.991		
aVARV-VK388	0.991	0.994	0.99	0.991	0.998	
aVARV-VK470	0.99	0.992	0.989	0.99	0.996	0.996

Table 4.2: Sequence similarity between VARV-VD21, aVARV-VK382, aVARV-VK388, and aVARV-VK470 consensus sequences and relevant orthopox reference sequences. The closest modern VARV was selected, excluding the two sequences recovered from Czech museum specimens published in Pajer *et al.* (2017) [115].

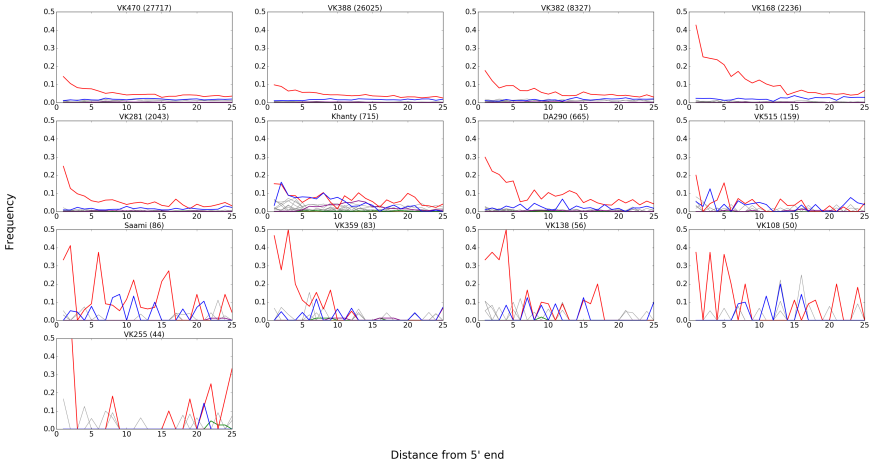


Figure 4.2: Ancient DNA damage patterns. The frequencies of the mismatches observed between the orthopoxvirus reference sequences and the reads are shown as a function of distance from the 5' end. C $\rightarrow$ T (5') and G $\rightarrow$ A (3') mutations are shown in red and blue, respectively. All other possible mismatches are shown in grey, and bases with soft-clipping are shown in orange. The number of reads matching the orthopoxvirus reference sequence for each sample is shown in parentheses.

Sequence	Suggested breakpoints	Detecting algorithm (P value)
aVARV-VK382	7928 (7624–8106) and 8189 (8110–8285)	3Seq (8.27×10^{-3}), BootScan (2.619×10^{-6}), GENECONV (3.683×10^{-4})
aVARV-VK470	1034 (undetermined) and 1942 (undetermined)	GENECONV (2.337×10^{-3})

Table 4.3: RDP4 recombination analysis. An RDP4 analysis found evidence of recombination in only small regions of three of the aVARV sequences. The suggested breakpoints offsets are for the ungapped sequences and indicate the most likely breakpoint location, followed by the 99% confidence interval of the breakpoint if determined. Detecting algorithms are followed by the average P value in parentheses. aVARV-VK388 was the suggested major parent in all cases, with an unknown minor parent. In all cases the RDP4 program issued the caveat that the proposed recombination signal may be attributable to a process other than recombination. No recombination signal was detected with aVARV-VK388 as the recipient sequence. No evidence was found to suggest that any other sequence in the selection is the product of recombination involving any aVARV sequence. *Table contributed by Terry Jones.*

To establish the phylogenetic placement of the three sequences from the high-coverage samples in relation to the overall diversity of the orthopoxviruses, we inferred an ML tree including sequences from the three high-coverage samples and 84 sequences chosen to represent the full orthopoxvirus diversity. The resulting tree (Figs. 4.9b, 4.3) shows that the Viking age sequences form a now-extinct monophyletic ancient VARV-like (aVARV) clade in a sister relationship with the clade consisting of all modern VARV and VARV-VD21 (collectively referred to as mVARV).

The evolutionary placement algorithm implemented in EPA-ng [284] can be used to place reads or partial sequences into a previously-inferred phylogenetic tree. Confirming the presence of the aVARV clade, the algorithm places seven of the eight Viking age low-coverage samples (VK108, VK168, VK255, VK281, VK359, VK515, and VK533) within the aVARV clade, and the other low-coverage Viking age sample (VK138) on the branch connecting the mVARV and aVARV clades. The two low-coverage samples dated to the 19$^{\text{th}}$ century (FIN1 and KHA1), are placed within the mVARV clade (Figs. 4.1b, 4.4).

A prior condition for making dated phylogenetic trees is the existence of a temporal signal in the sequence data. A linear regression of root-to-tip distances from an ML tree that includes sequences from the high-coverage samples and mVARV shows clear evidence of clock-like evolution (Fig. 4.5), supporting earlier results [116,277]. Consequently, we proceeded to infer dated coalescent trees using BEAST2 [150] under six different evolutionary models (Table 4.4). Under the best-fitting model, the MRCA of the combined mVARV and aVARV sequences is dated to 1.6 thousand years ago (kya) (95% highest priority density interval (HPD95): 2.1–1.4 kya). The age of the aVARV clade is estimated to be 1.5 kya (HPD95: 1.7–1.3 kya). The age of the MRCA of the mVARV clade is dated to 0.43 kya (HPD95: 0.47–0.35 kya), and 0.3 kya (HPD95: 0.42–0.2 kya) for the mVARV clade without VARV-VD21, both of which are in line with previous molecular dating analyses [115, 116, 277, 278] (Table 4.5, Fig. 4.6).

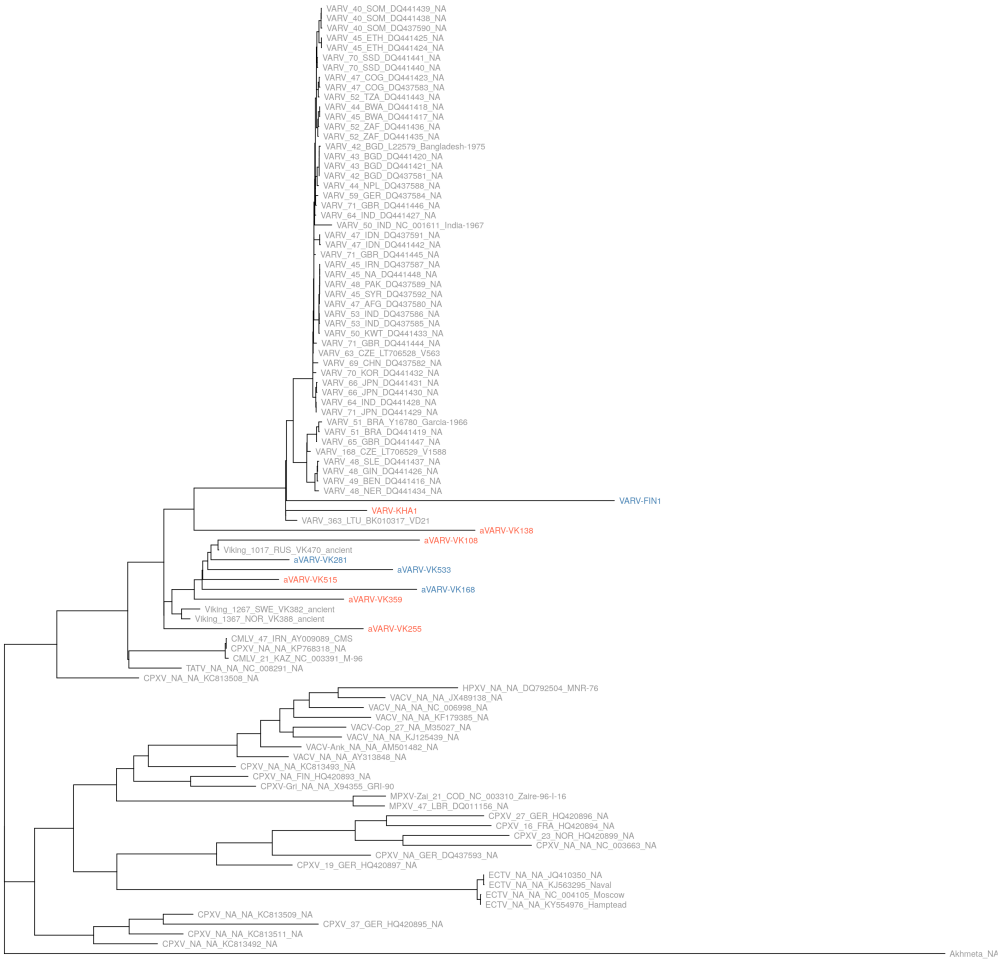


Figure 4.4: Best placements of low-coverage samples. The maximum likelihood tree from Fig. 4.3 showing the most likely placements of confirmed (blue) and likely (red) samples. Terminal branch lengths of low-coverage samples are likely elongated because of increased inaccuracies in the consensus sequences, due to low coverage. The x-axis denotes substitutions per site (s/s). The tree is rooted with Akhmeta_NA_NA_MH607143_Vani_2010. *Figure contributed by Martin Sikora*.

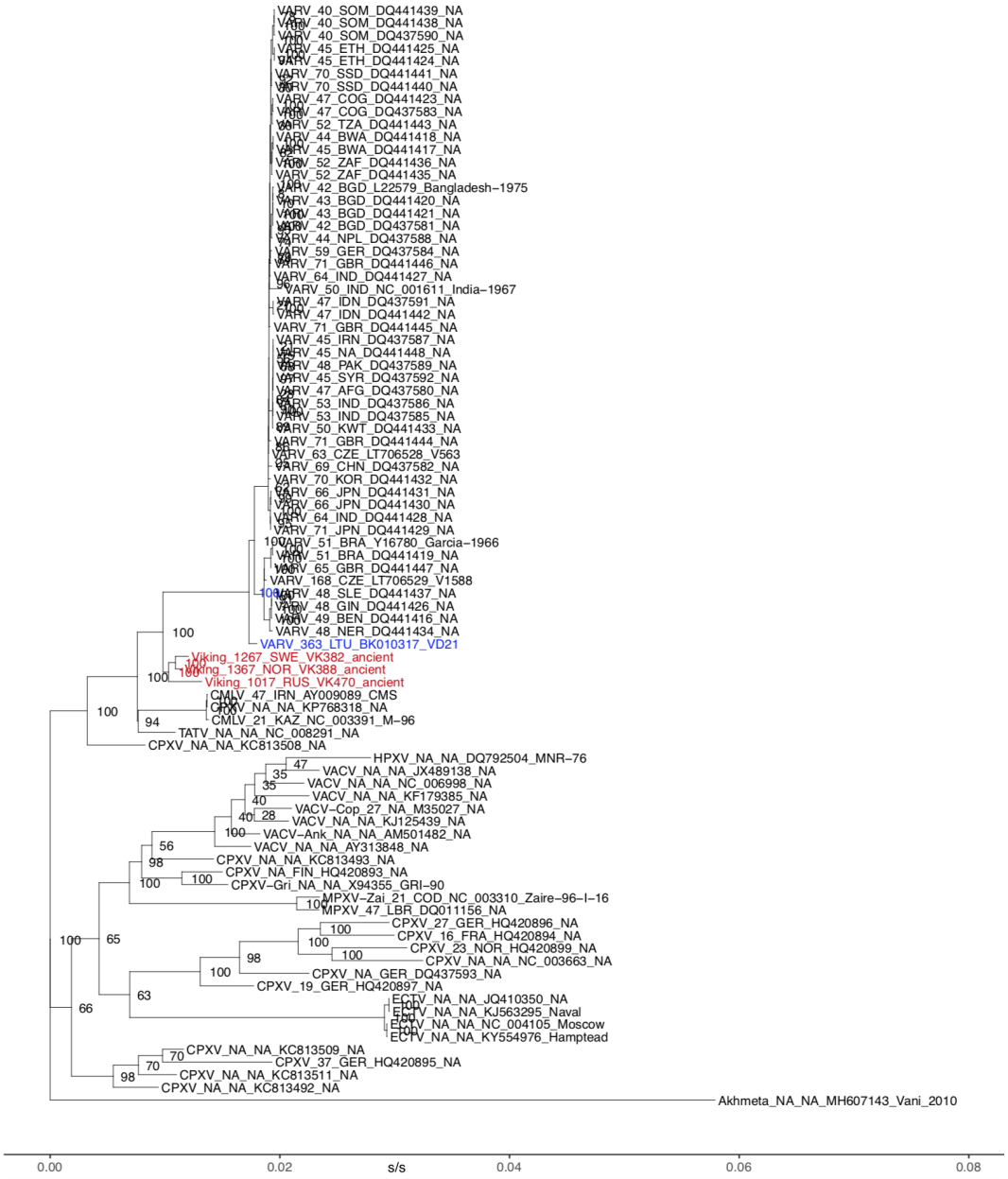


Figure 4.3: Maximum likelihood tree of the orthopoxviruses and the Viking age high-coverage samples used for EPA analysis. Sequences were aligned using MAFFT. The tree was inferred using RAxML-ng with a K81uf+G+I substitution model, and an ML estimate of tree topology, branch lengths, substitution rates, and nucleotide frequencies and support estimated using 1000 bootstrap replicates. Ancient sequences from this publication are shown in red, the ancient sequence from Duggan *et al.*, (2016) in blue [116]. Taxon names: virus/historical period, sample age relative to 2017, country abbreviation of sequence origin and region of sequence origin, as determined by the Standard country or area codes for statistical use, accession number/sample identifier, additional remarks. The x-axis denotes substitutions per site (s/s). The tree is rooted with Akhmeta_NA_NA_MH607143_Vani_2010.

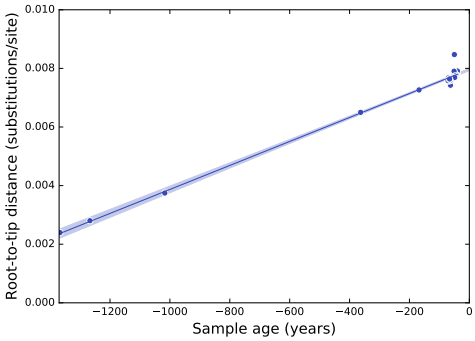


Figure 4.5: Linear regression of root-to-tip distances against sampling dates. Root-to-tip distances were extracted from an ML tree including the three high-coverage samples and mVARV sequences using TempEst [225], with the 'best-fitting-root' option. The regression analysis was performed in SciPy [226]. Shaded areas represent 95% confidence. Mean sample ages were used. The slope is 4.09×10^{-6} and the correlation coefficient is 0.99 (R^2=0.99).

Model (molecular clock, population prior)		Strict, coalescent constant	Strict, coalescent exponential	Strict, Bayesian skyline	Log-normal relaxed, coalescent constant	Log-normal relaxed, coalescent exponential	Log-normal relaxed, Bayesian skyline
	Likelihood	-152877.65	-152876.67	-152870.85	-152807.89	-152807.42	-152801.48
Strict, coalescent constant	-152877.65	0	0.98	6.8	69.76	70.23	76.17
Strict, coalescent exponential	-152876.67	-0.98	0	5.82	68.78	69.25	75.19
Strict, Bayesian skyline	-152870.85	-6.8	-5.82	0	62.96	63.43	69.37
Log-normal relaxed, coalescent constant	-152807.89	-69.76	-68.78	-62.96	0	0.47	6.41
Log-normal relaxed, coalescent exponential	-152807.42	-70.23	-69.25	-63.43	-0.47	0	5.94
Log-normal relaxed, Bayesian skyline	-152801.48	-76.17	-75.19	-69.37	-6.41	-5.94	0

Table 4.4: Model testing for different clock models and population priors. Models were compared using path sampling, as implemented in BEAST2. Log marginal likelihood values ('likelihood') were compared using a Bayes factor test. A positive value for the Bayes factor implies support for the column model, a negative value support for the row model. According to Kass and Raftery, a Bayes factor in the range of 3–20 implies positive support, 20–150 strong support, and >150 overwhelming support [229].

Model (molecular clock, population prior)	Median root ages								Substitution rate (s/s/y)
	Root		aVARV clade		mVARV clade		VARV clade		
Strict, coalescent constant	1692	(1556, 1841)	1520	(1443, 1596)	433	(403, 470)	330 (297, 368)		$4.99x10^{-6}$ ($4.36x10^{-6}$, $5.61x10^{-6}$)
Strict, coalescent exponential	1697	(1572, 1847)	1522	(1448, 1603)	435	(404, 471)	335 (303, 370)		$4.97x10^{-6}$ ($4.34x10^{-6}$, $5.62x10^{-6}$)
Strict, Bayesian skyline	1697	(1566, 1845)	1522	(1446, 1599)	435	(403, 471)	336 (302, 371		$4.98x10^{-6}$ ($4.36x10^{-6}$, $5.65x10^{-6}$)
Log-normal relaxed, coalescent constant	1578	(1349, 1957)	1451	(1323, 1607)	424	(349, 556)	303 (203, 408)		$5.54x10^{-6}$ ($4.12x10^{-6}$, $6.94x10^{-6}$)
Log-normal relaxed, coalescent exponential	1724	(1391, 2301)	1513	(1348, 1731)	430	(348, 556)	303 (209, 412)		$5.05x10^{-6}$ ($3.49x10^{-6}$, $6.55x10^{-6}$)
Log-normal relaxed, Bayesian skyline	1636	(1364, 2138)	1472	(1330, 1664)	434	(346, 570)	304 (203, 423)		$5.37x10^{-6}$ ($3.9x10^{-6}$, $6.91x10^{-6}$)
Previously published									
Duggan *et al.*, 2016 [116], Strict, coalescent constant					372–429		224–283		$8.5x10^{-6}$ ($7.3x10^{-6}$, $9.6x10^{-6}$)
Smithson *et al.*, 2017 [278]					454–568		345–457		No rate published
Pajer *et al.*, 2017 [115]					667		322		No rate published
Porter *et al.*, 2017 [277]							372–463 / 375–419		$5.44x10^{-6}$ ($4.73x10^{-6}$, $6.16x10^{-6}$) / $8.27x10^{-6}$ ($7.48x10^{-6}$, $9.10x10^{-6}$)

Table 4.5: Median root ages and substitution rates inferred using BEAST2 under different clock models and population priors, compared to previously published estimates. Numbers in parentheses indicate 95% highest posterior density intervals. A TPM1 substitution model with unequal base frequencies, invariant sites, and gamma distributed rate heterogeneity among sites was used throughout.

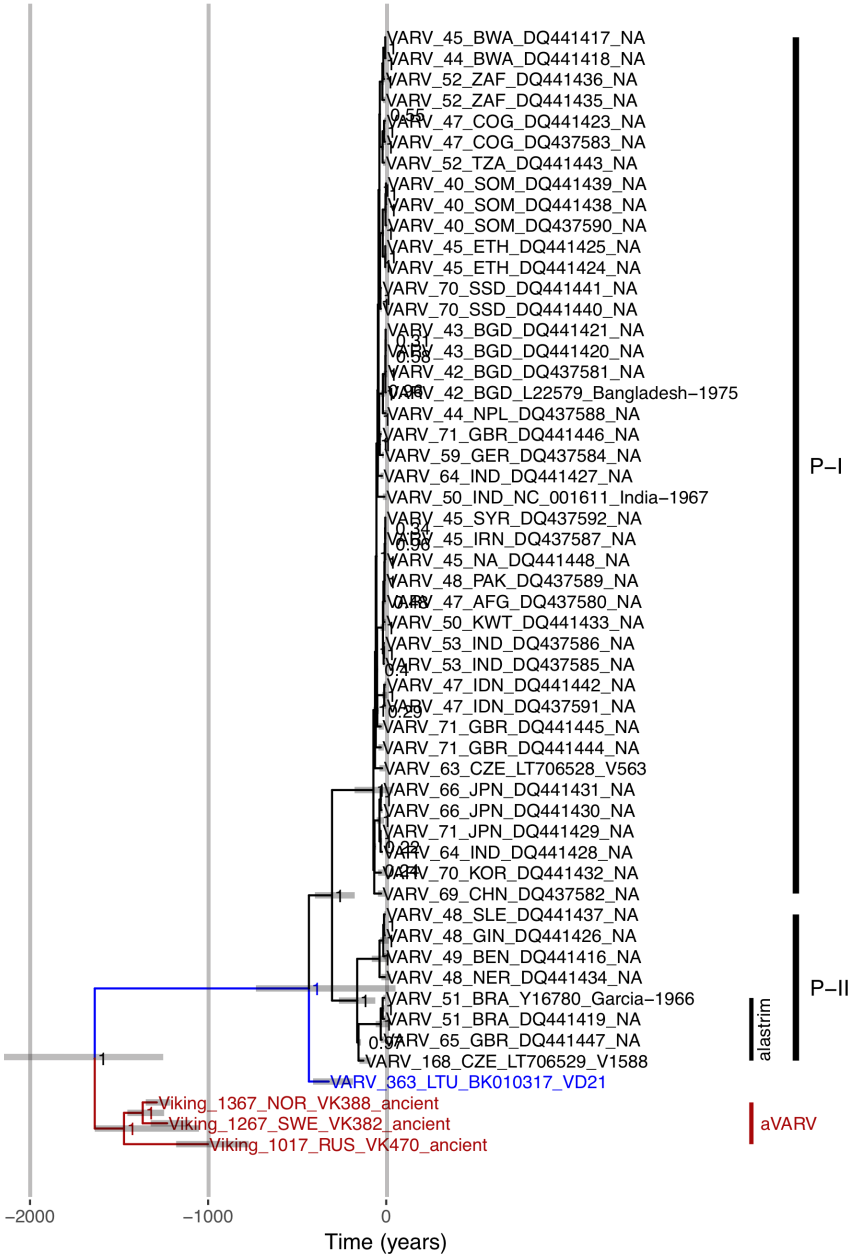


Figure 4.6: Maximum clade credibility tree inferred with BEAST2. A log-normal relaxed clock and coalescent Bayesian skyline population prior are used, as well as a TPM1 substitution model with unequal base frequencies, invariant sites, and gamma distributed rate heterogeneity among sites. The substitution rate is inferred as 5.4×10^{-6} s/s/y (95% HPD interval: 3.9×10^{-6} to 6.9×10^{-6} s/s/y), and the root age as 1,636 years old (yo) (95% HPD interval: 1,364 to 2,140 yo). The x-axis denotes time into the past, in years. aVARV sequences in this paper are shown in red, the ancient sequence from Duggan *et al.*, 2016 [116] in blue. Taxon names: virus/historical period, sample age relative to 2017, country abbreviation of sequence origin, accession number/sample identifier, additional remarks. Horizontal grey bars indicate 95% HPD intervals of node ages. P-I and P-II refer to primary clade I and primary clade II, respectively [289].

Several genes are inactivated or lost in mVARV viruses compared to other orthopoxviruses, possibly due to host adaptation [270, 274]. We therefore compared and contrasted the gene content of sequences from the three high-coverage aVARV samples, mVARV, plus CMLV and TATV (the orthopoxviruses most closely related to aVARV) (Figs. 4.7a–d). We find a gradual inactivation of genes over time (Fig. 4.8). Loss of host-range genes has been correlated with an increase in host specificity [273], suggesting that the aVARVs may have had a broader host range than mVARV. However, host range in poxviruses is likely to be a multigenic trait depending on the cooperation of several genes, as is the case in myxoma virus [290] and VACV [291–295]. The differential pattern of gene inactivation thus does not necessarily imply that the aVARVs had a host tropism for animals other than humans, although this is a possibility.

We next considered the presence and inactivation of specific genes (Fig. 4.9), by first examining the pattern within the aVARV clade, and then comparing that to the pattern in the mVARV clade.

Within just the aVARV clade, we observe a diverse pattern of gene inactivation, (Figs. 4.7a–d), 4.9) among 14 genes that are inactivated in some, but not all, of the sequences from high-coverage samples. These results show that the overall consistent pattern of gene loss and inactivation in mVARV was preceded by a period during which there were co-circulating VARV strains with a diverse pattern of gene inactivation. Diversity within the aVARV clade is further supported by the phylogenetic placement of the low-coverage samples (Fig. 4.4). Geographic dispersion and mobility in the Viking Age with subsequent isolation, may have led to the creation and co-circulation of the diverse lineages in the aVARV clade. The uniformity in the mVARV clade may reflect the eventual dominance of one particular strain. Before that state was reached, and because different subsets of genes can evidently be lost independently, the virus appears to have performed a simultaneous parallel exploration of the space of possible combinations of gene inactivation, as shown in the aVARV clade.

(a)

(b)

(c)

(d)

Figure 4.7: Similarity and presence and inactivation of genes in the ancient samples. a) VARV-VD21. b) aVARV-VK382. c) aVARV-VK388. d) aVARV-VK470. Genes are shown on the x-axis. The figures are split in two parts. The bottom part shows the presence (filled in circle), absence (empty circle) and inactivation (plus sign) of genes in black in the 17 orthopoxvirus reference sequences from Dataset 3. The top part shows the similarity of the reads from an ancient sample to the modern gene sequence. If an ancient sample had reads covering more than 50% of the reference gene sequence, and the median bit score was within 2 of the highest median bit score for a particular gene, a coloured dot is shown in the row of the modern virus in question. The row labelled 'Highest' indicates the reference sequence with the highest bit score for each particular protein. A black square indicates that the modern reference sequence of a gene is inactive. A plus sign indicates a gene-inactivating mutation in the ancient consensus sequence, and a minus sign indicates an uncertain gene-inactivating mutation in the ancient consensus sequence. Black triangles in the top two rows indicate known host range [273] and diploid genes [270], respectively. Five genes on either side of the genome are greyed out and were not considered in other analyses, since they are diploid and identical. Gene labels correspond to the end offset of the gene in CPXV-Gri/GER, the name of the VACV-COP homolog, if applicable, separated by '|'. A figure with better resolution can be found at `http://antigenic-cartography.org/barbara/phd-thesis/`.

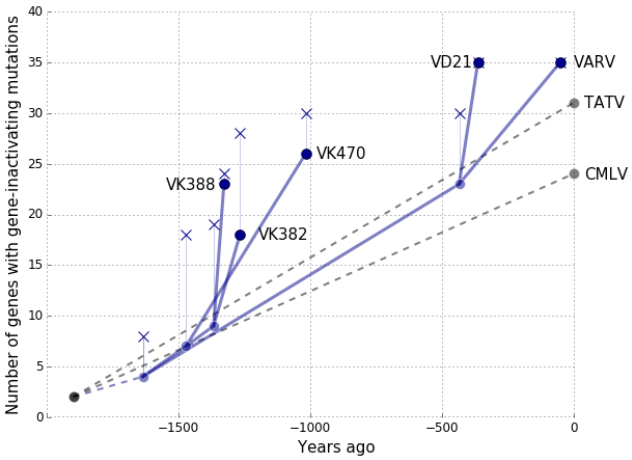


Figure 4.8: Gene inactivation over time. The number of genes with gene-inactivating mutations is shown on the y-axis, plotted against time (in years before present) on the x-axis. The evolutionary relationships between the nodes and tips (as given by the ML tree in Fig. 4.1) are indicated by solid lines. Dates for internal nodes are taken from the dated coalescent tree in Fig. 4.6, estimated using a relaxed log-normal clock and a Bayesian skyline population prior. Gene counts for internal nodes are inferred based on identical gene-inactivating mutations in their descendant viruses. Note that no age has been inferred for the node ancestral to CMLV, TATV, and the ancestor of mVARV and aVARV, because those viruses evolved in different hosts. Therefore, the root node is shown on the left-hand side of the figure at an arbitrary time point, and is connected by dotted lines to indicate the uncertain timing of gene inactivations en route to those three descendant nodes. Blue dots indicate counts of gene-inactivating mutations that can be determined with certainty (including genes that are absent in VARV and that have no coverage in the ancient samples). The counts for the data points plotted with an 'x', vertically above tips or internal nodes, also include gene-inactivating mutations that cannot be identified with confidence and genes with no coverage in the ancient samples when they have coverage in VARV.

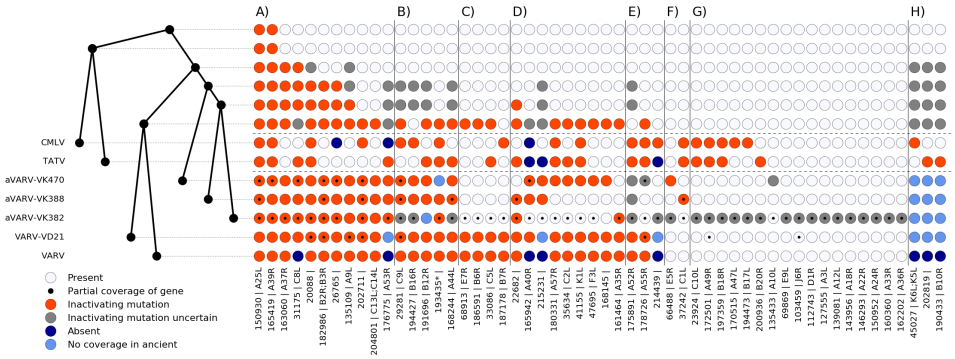


Figure 4.9: Gene-inactivating mutations. The cladogram on the left shows the topology of the phylogenetic relationship between the mVARV, the aVARV sequences we report in this paper, and TATV and CMLV (see trees in Figs. 4.1b), and 4.6). Each row to the right corresponds to the virus at that level in the cladogram, including rows with inferred gene state for internal nodes (top six rows). The 54 columns represent genes that are either absent or that have an inactivating mutation in at least one of the virus rows. Genes with inactivating mutations for a particular virus or node are shown in red, and those that cannot be identified with certainty (Materials and Methods) in grey. Genes that are absent in a modern virus are dark blue, and those with no coverage in VARV-VD21, and the aVARV sequences are light blue. In cases where full coverage might not be available (VARV-VD21 and the aVARV sequences), genes with partial coverage are marked with a black dot. Gene labels indicate the final offset of the gene in CPXV-Gri/GER and the name of the VACV-COP homolog [270], if applicable, separated by 'l'. Genes are grouped into eight categories: A) Inactivated in mVARV and aVARV sequences. B) Inactivated in mVARV and uncertain inactivations in some but not all aVARV sequences. C) Inactivated in mVARV, but present in aVARV sequences. D) Inactivated in mVARV, but present in some but not all aVARV sequences. E) Inactivated in mVARV, and uncertain inactivations in aVARV sequences. F) Present in mVARV, but absent in some aVARV sequences. G) Present in mVARV, but with uncertain inactivations in some aVARV sequences. H) Absent or no coverage in mVARV and aVARV sequences, and with gene-inactivating mutations in CMLV or TATV.

To focus on the differences between the mVARV and aVARV clades, we organized gene presence or inactivation status into eight categories, A–H (Fig. 4.9, and Appendix A). We found three groups of these categories of particular interest: genes inactive in both clades (categories A and B), genes inactive in mVARV but active in at least one aVARV sequence (categories C, D, and E), and genes active in mVARV but inactivated in at least one aVARV sequence (category F).

In the first group of interest (genes inactive in both clades), the exact position of mutations in genes that are inactivated in both mVARV and aVARV allows us to infer whether intact genes were likely present in the common ancestors of these viruses (Fig. 4.9, Appendix A). Of the 16 genes that are inactivated in both mVARV and

aVARV, five (*A25L, A37R, A39R, C8L,* and possibly 20088) have an identical inactivating mutation in the mVARVs and aVARVs (category A and B in Fig. 4.9, Appendix A). This indicates a single introduction of the ancestor of mVARV and aVARV into humans, and the continuous evolution of VARV in humans for at least the last ~1600 years. This is as opposed to multiple independent introductions with subsequent gene-inactivating mutations in identical positions in all five genes. The other 11 genes that are inactivated in both clades, but in different ways, suggest a degree of parallel evolution within the human host after the divergence of the mVARV and aVARV clades.

In addition, a gene inactivated in both mVARV and aVARV may have induced a similar outcome after infection of humans. The VACV gene encoding a soluble IL-1β receptor is inactive in VACV strain Copenhagen (VACV-COP) (gene *B16R*) but functional in VACV strain Western Reserve (VACV-WR) where it is called gene *B15R* [296]. Infection of mice with either VACV-COP or VACV-WR in which the gene *B15R* was inactivated, resulted in a febrile response [276]. Conversely, infection with either virus expressing a functional IL-1β receptor prevented the induction of fever [276]. Human infections with VARV, which all have a disrupted *B16R* gene, are accompanied by high fever [10]. The absence of *B16R* in the aVARV clade may suggest that disease caused by those viruses also included high fever.

In the second group of interest (categories C, D, and E in Fig. 4.9), 16 genes are inactivated in mVARV but are present or have an undetermined status in some or all of the aVARV sequences. Nine of these genes (*B7R, A40R, C2L, K1L, F3L, A35R, A52R, A55R*, and *CrmE*) encode known virulence factors or immunomodulators in VACV, three of which are members of the kelch-like family (*C2L, A55R*, and *F3L*, Appendix A) [7]. Deletion of the kelch-like proteins *C2L* or *A55R* in VACV-WR leads to increased lesion size in intradermally infected mice [297, 298], while the virulence of VACV lacking *F3L* is reduced in the same model [299]. Deletion of kelch-like proteins has also been proposed to reduce the host range and virulence of CPXV-GRI90 in vitro [300]. Also in contrast with mVARV, aVARV-VK382 and aVARV-VK388 encode a predicted functional guanylate kinase (*A57R*). This gene is otherwise only active in CPXV, and appears to have been lost independently in CMLV, TATV, mVARV, and aVARV-VK470 (Appendix A), suggesting that the gene was functional in their common ancestor.

Finally, in the third group (category F in Fig. 4.9), two genes (*E5R* and *C1L*) are inactivated in at least one aVARV but in none of the mVARV sequences. The effect of the inactivation of *E5R* and *C1L* in VACV is unknown.

Hypotheses regarding the earliest history of VARV have been derived exclusively from often ambiguous historical accounts, and the visual examination of mummies dating from as early as 3570 ya [279, 280]. The aVARV sequences push the date of the oldest definitive VARV infections back by ~1000 years. The dates of the earliest likely presence of smallpox in specific regions are difficult to determine from written records and have in some cases been a matter of dispute [10, 279, 280, 301]. Ancient DNA sequences can resolve such questions by providing clear evidence of VARV presence in time and space. The dating of the aVARV samples, from as early as 550 CE, matches that of multiple written accounts of likely smallpox infections in southern and western Europe from the late 6th century onwards [10, 279, 280, 301–303]. Our finding of the virus in northern Europe at these times disproves various suggestions of first introductions involving later dates. For example, the introduction (or later establishment) of smallpox into Europe via returning Crusaders [10, 272], into Europe via Spain during the Moorish invasion of 710 CE [272], or into England by Norman invaders in 1241–2 CE [304], as well as the claim that the virus had not reached northern Europe by 1000 CE [10]. The written record, together with the confirmed aVARV infections, suggest a pan-European presence of the virus by the end of the Viking Age at the latest, as has been suggested for parts of southern and western Europe based purely on written records [279, 280, 302].

We found evidence of VARV in 11 of 555 Viking age individuals (~2%). If aVARV, like mVARV [10], did not lead to persistent infections and if no viral particles remained in teeth or bone following an acute infection, these individuals all died with an active infection. The 11 samples we identified as positive were likely not the only infections however, because the presence of viral DNA in ancient teeth and bones is almost certainly an imperfect diagnostic test of viral infection and because VARV is not thought to be viremic for the entire duration of the illness [10]. We can consider what this 2% figure might indicate, depending on the character of aVARV infections in northern Europe during the Viking Age. If the infections were very rarely fatal, as with modern variola minor [10, 280], the rate would represent an overall approximate prevalence of VARV. At the other extreme, if the disease was often fatal, as observed for variola major [10, 280], apart from two individuals who died of other causes, it is possible that all nine others died as a result of their infections. In that case, the rate might be indicative of the overall percentage of deaths due to smallpox. Both numbers would be an underestimate, due to the imperfect diagnostic test. By comparison, the London Bills of Mortality show a percentage of deaths due to smallpox as ~2% in 1650 CE [305].

The Viking age VARV samples reveal a now-extinct virus clade with a diverse pattern of gene inactivation that differs markedly from modern viruses. Following a single introduction ~1.6 kya, VARV followed at least three parallel evolutionary paths. This surprising new snapshot of the evolution of the virus comes from a restricted geographic and temporal range. Given the long global spread of smallpox, it can be expected that additional ancient sequences will further increase our knowledge of the diversity and evolution of VARV.

CHAPTER 5: USING ANCIENT DNA TO STUDY VIRUS EVOLUTION: DISCUSSION AND CONCLUSION

In the previous three chapters, I showed that it is possible to recover hepatitis B virus (HBV), human parvovirus B19 (B19V), and variola virus (VARV) genomes from human DNA samples up to ~4,500, ~6,900, and ~1,400 years old, respectively. Another group has concurrently showed the same for a ~7,000 year old HBV sequence [112]. The previous chapters have presented the three largest datasets of ancient viral sequences described to date. In the absence of ancient sequences, our understanding of important aspects of viral evolution has been based on extrapolations from modern data. Upon inclusion of the ancient sequences, we find a complexity of virus evolution that had not been appreciated when only modern sequences were considered:

- The close sequence similarity of some of the ancient HBV and B19V sequences to sequences of genotypes still circulating today is surprising, given the fast substitution rates observed when considering modern sequences of those viruses [37,126,157,158][1]. In particular, nine out of twelve ancient HBV sequences, the oldest of which is ~4,300 years old, and all ten available B19V sequences (up to ~6,900 years old), can be assigned to modern genotypes, according to the genotype definition criteria used for modern sequences[2].

- At the same time, the studies on ancient viral sequences revealed variants of the virus that are now extinct, such as the high-coverage ancient VARV sequences, or the ancient HBV sequences most closely related to modern non-human primate HBVs. This shows that virus diversity existed in the past which we have no knowledge of when only considering modern sequences, and which could change current interpretations about the origins of modern variants of those viruses.

- Furthermore, dated coalescent trees inferred using modern and the ancient sequences show that for HBV and B19V, extrapolations of most recent common

[1] We observe between 92.5% to 98.7% and 95.2% to 98.2% sequence similarity to modern sequences, for HBV and B19V, respectively.

[2] For HBV, sequence differences within human genotypes are up to 7.5%, and up to 4.5% within subgenotypes [193]. Furthermore, modern genotypes and subgenotypes are informally characterised by sequence length, insertions and deletions, serotype, preferred transmission route, and pathogenicity [193, 242]. Non-human primate sequences are not divided into genotypes. For B19V, genotypes are defined based on sequence identity, with sequence difference ~10% within genotype 1, and ~5% within genotypes 2 and 3 [243, 255].

ancestor dates and substitution rates are very sensitive to the inclusion of the ancient sequences. The results also call into question some of the current thinking about substitution rates (see chapter 5.1, below).

– The ancient sequences provide spatio-temporal reference points during the evolution of particular viruses. This allowed us to revise narratives about timings, origins, and first introductions of virus variants in certain geographic areas, such as the postulated origin of HBV genotype A in Africa, or the presence of the ancestor of modern VARV in Europe.

– Finally, the ancient sequences provide a catalogue of variation that has existed in the past, which may highlight features of the virus that are more prone to change in the future. These include the genes lost along the evolutionary path to modern VARV, or the absence of the typical six nucleotide insertion for HBV genotype A in the three oldest genotype A sequences. The phenotypic characterisation of some of this variation may improve our understanding of its fitness effects, the pathogenicity of the virus in the past, and the potential of the virus to change in the future. This is exemplified by the work on the reconstructed influenza A/H1N1 virus from the 1918 pandemic, which highlighted an increased lethality in mice and suggested the presence of a novel cleavage mechanism of the hemagglutinin [168]. Furthermore, studying the efficiency of vaccines, antivirals, and diagnostic tests on the ancient viruses may provide information about their utility in the future.

Even though ancient viral sequences have improved our understanding of the evolution of HBV, B19V, and VARV, a note of caution is appropriate at this stage. The ancient sequences currently available (twelve, ten, and three (and an additional ten low coverage) ancient sequences, for HBV, B19V and VARV, respectively) are a tiny number over a very large temporal and spatial range, and it would be premature to assume that they represent the final story about the evolution of HBV, B19V, and VARV. Further ancient sequences will most likely expand and refine some of the conclusions drawn in chapters 2 – 4, in particular about the spatio-temporal distribution of genotypes, past genetic diversity, and divergence dates. While we have taken great care to not make claims that are wrong due to the small amount of data that was available to us, and make appropriate caveats, additional ancient sequences may not always agree with the results presented in the previous chapters.

While studying the ancient sequences for each pathogen on its own offers valuable insights into its evolution, I next expand on one overarching topic from the work on both HBV and B19V, concerning the low substitution rates inferred for those viruses

when the ancient sequences are included, as opposed to the high rates when only modern sequences are considered.

5.1 SUBSTITUTION RATES

The substitution rates we estimated using the ancient HBV and B19V sequences and a set of modern sequences are about an order of magnitude lower than the rates estimated using modern sequences alone. The decrease of substitution rates inferred over longer timescales using the ancient sequences is in agreement with observations that others have made: rates estimated using external calibrations such as dated fossils or biogeographic events that can be millions of years old, are lower than rates estimated using modern heterochronously sampled sequences, which may only span a timescale of a few weeks to decades [37, 152, 161, 162]. This phenomenon is called the 'time dependent rate phenomenon' [161]. For viral substitution rates, it has been shown that the process of rate decay with increasing sampling interval is gradual and can be described with a power-law relationship [152]. Rate decay is similar for different Baltimore groups [152]. Using the ancient sequences, we can also observe a somewhat gradual rate decay in the HBV and B19V datasets when subsequently adding older sequences, although rates inferred from modern sequences alone are higher, possibly due to the absence of temporal signal [113] (Fig. 5.1).

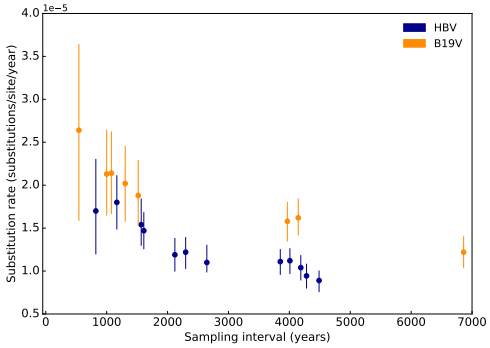


Figure 5.1: Decrease of substitution rate with increasing sampling interval in the HBV and B19V datasets. Rates were inferred under a strict clock and coalescent Bayesian skyline population prior for datasets generated by successively adding the next-oldest sequence (sequences of the same age were added together). Rates computed from modern sequences alone were omitted, due to the suspected lack of temporal signal. Modern rates were: HBV: $4.0\text{x}10^{-5}$, (95% HPD interval: $4.8\text{x}10^{-6}$, $1.0\text{x}10^{-4}$) s/s/y, B19V: $1.39\text{x}10^{-4}$, (95% HPD interval: $9.01\text{x}10^{-5}$, $1.86\text{x}10^{-4}$) s/s/y.

Interestingly, VARV does not show the same downward trend of the substitution rate as HBV and B19V when older ancient sequences are included. The rate we find ($5.4\text{x}10^{-6}$ s/s/y (95% HPD interval: $3.9\text{x}10^{-6}$ to $6.9\text{x}10^{-6}$ s/s/y)) is only slightly lower than the rates Duggan *et al.,* (2016) inferred using similar methods to us (BEAST 1.7

[306] as opposed to BEAST2), both with and without the ~360 year old Lithuanian mummy sequence [116]. This may be caused by the already strong temporal signal in the modern data alone, the relatively young age of the oldest ancient VARV sequence (~1,400 years) compared to HBV (~4,500 years) and B19V (~6,900 years), as well as by the tree topology, where the ancient sequences fall basal to the modern sequences in the tree, as opposed to within or between existing clades.

Ho *et al.,* (2011) discuss possible causes of the discrepancies in substitution rate observed over short and long timescales [161]. They include that substitution rates observed over short timescales are inflated due to the presence of sequencing errors, transient mutations, and calibration errors (such as assuming that the time of population divergence and genetic divergence correspond). Substitution rates inferred over long timescales may be underestimated due to substitution model inaccuracy and inaccuracy when modelling rate heterogeneity among sites, resulting in underestimation of saturation. Furthermore, changes in virus biology and selection pressure over time may also lead to changes to the substitution rate over different timescales [161]. At least two of those possible causes are now considered unlikely: Since rate decay is gradual over time and comparable between different Baltimore groups [152, 162], explanations that would result in discrete changes in substitution rate, or that would affect viruses in different Baltimore groups differently, such as short-timescale changes in virus biology or selection pressure, seem unlikely [38, 152, 307]. Also, deleterious mutations may be removed quickly from the population, which would mean that they have little effect on long term substitution rates [307].

The datasets used in this thesis have the widest time span ever used for molecular dating analyses of viruses based on heterochronous data. At the same time, the viral diversity in these datasets is a severe under-sampling of the actual viral diversity present during the time span the datasets cover. This is especially pronounced at deeper time scales, since the modern viral diversity is more extensively sampled (though most likely still incomplete). While the ancient sequences provide important calibration points for molecular dating analyses, the under-sampling of the viral diversity, especially at deeper timescales, likely impacts the result of such analyses. I would like to focus on two areas that are affected by this. Firstly, under-sampling can lead to difficulties in determining which mutations are fixed, and secondly, it may exacerbate problems with adequately modelling the substitution process when performing molecular dating analyses.

The substitution rate is defined as the rate at which mutations become fixed in the population, measured as substitutions per site in the genome per time interval [35]. 'Fixation' refers to the process where at a particular site in the genome, one nucleotide or

amino acid completely replaces another within a population [308]. The rate of fixation depends on the population size and the fitness effect of the mutation [308]. In a dataset that is well-sampled in space and time, informally determining which mutations are becoming fixed is relatively straightforward, via identifying positions with changing frequencies of nucleotides or amino acids. However, in the context of studies on datasets that include sparsely sampled viral sequences thousands of years old, there are two problems related to the word 'fixed' in the definition of the substitution rate. The first is that as a result of the severe under-sampling of the viral diversity in such datasets, we cannot know whether observed mutations are present in enough individuals to be considered fixed. The second problem, more subtle, is that the word 'fixed' in the definition of the substitution rate contains an unacknowledged implicit time component. In the almost ubiquitous 'substitutions per site per year', the 'year' unit of time is rarely considered, and has been an appropriate and useful yardstick for datasets consisting of modern sequences only. Unless a nucleotide at a particular site arises and never changes to another nucleotide, a substitution is only fixed over a certain time interval (1 day, 10 years, 100 years etc.) before being replaced by another. But in the context of datasets including sequences thousands of years old, substitutions that may look fixed over a timescale of a single year, may appear as transient mutations over a much longer timescale. When looking at a set of sequences sampled over a short time interval, many more substitutions appear as fixed, even though in the long term they may change, or the lineages may go extinct. These two problems mean that in the context of current ancient DNA studies of viral populations, there is no clear understanding of what constitutes a fixed substitution, and therefore it is also unclear which substitutions should be relevant in the calculation of the substitution rate.

Models of sequence evolution are theoretically able to correct for multiple changes at the same site [308]. Estimates of divergence dates with closely-related modern sequences sampled over short timescales correspond well with epidemiological evidence of those divergence events. For example, by estimating the most recent common ancestor of swine and human segments of the pandemic 1918 H1N1 virus, molecular dating analyses suggest the emergence of the virus shortly before 1918 [166]. Likewise, the divergence date of HIV group M corresponds well with the growth of major cities in the area where ancestral forms of the virus can still be found in wild chimpanzee communities [309]. However, current models of sequence evolution may not be able to adequately correct for multiple changes at the same site over larger timescales and large evolutionary distances [310], which may lead to lower substitution rates inferred over large timescales. For example, Wertheim and Kosakovsky Pond (2011) observed that purifying selection may lead to artifi-

cially young divergence dates for modern heterogenously sampled pathogens [311]. Furthermore, using simulations, Soubrier *et al.,* (2012) showed that not sufficiently taking into account rate heterogeneity among sites leads to the underestimation of substitution rates in a time-dependent manner [312].

Using the ancient sequences, I can show that substitution rates may not only depend on the time scale of measurement, but also on the number of sequences of intermediate age between the youngest and the oldest sequence included in the dataset. For B19V, the substitution rate estimated with modern sequences and just the oldest ancient sequence (DA251) is lower than the substitution rate estimated from the same data but including the other nine ancient sequences of intermediate age (Fig. 5.2).

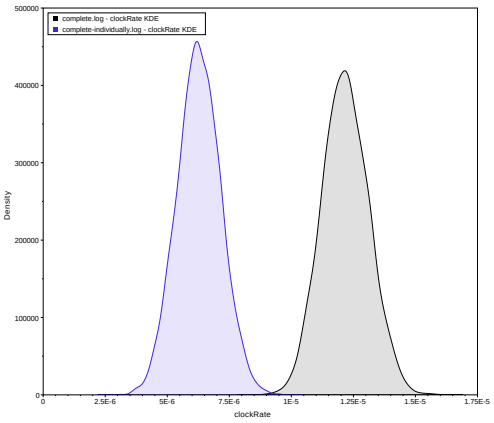


Figure 5.2: Kernel density estimation of the substitution rate for B19V inferred using all 10 ancient and the modern sequences (complete, black), or only the modern sequences and the oldest (DA251) ancient sequence (complete-individually, blue). The substitution rate estimated using just DA251 and the modern sequences is lower than the substitution rate estimated using all 10 ancient and the modern sequences. The substitution rate was inferred in BEAST using a strict clock and coalescent Bayesian skyline population prior.

Furthermore, the substitution rate inferred for B19V genotype 1 and the oldest ancient sequence (DA251) increases gradually as more ancient sequences with ages between DA51 and the modern sequences are included (in this case samples NEO105, DA66, and DA337) (Fig. 5.3). This pattern of rate decay as fewer intermediate sequences are included may be additional evidence that the substitution model used is unable to accurately model the substitution process when sequences of intermediate age are not present in the data. Therefore, the fact that dating analyses incorporating ancient sequences severely under-sample the viral diversity at deeper timescales may lead to artificially lower substitution rates.

If we do have a model of sequence evolution that correctly accounts for all changes that have taken place over the evolutionary history of a virus, or if we had access to all sequences of a virus that ever existed, what would the substitution rate be? Given

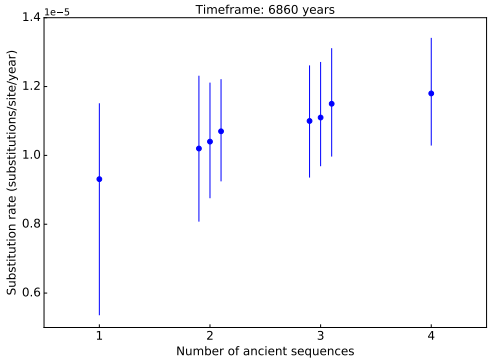


Figure 5.3: Substitution rates inferred for B19V genotype 1 and the oldest ancient sequence (DA251, 6862 years old (yo)), with different number of intermediate sequences also included. The substitution rate was inferred in BEAST using a strict clock and coalescent Bayesian skyline population prior. The x-axis indicates the number of ancient sequences included with the modern genotype 1 sequences: 1: DA251 only, 2: DA251 and either NEO105 (544 yo), DA66 (1,518 yo), or DA337 (3,967 yo). DA336 was excluded due to its similar age to DA337 (4146 yo) and its lower coverage. 3: DA251 and either NEO105 and DA55, NEO105 and DA337, and DA66 and DA337. 4: NEO105, DA66, DA337, DA251.

the results in Fig. 5.3, one may argue that the rate would be closer to that observed when we estimate rates using just the modern sequences.

While the 'time dependent rate phenomenon' can be caused by a number of factors, substitution rates will always depend on the timeframe of measurement, unless we have access to a subset of sequences representative of all changes that have taken place over the evolutionary history of a particular virus (as indicated by the often reasonable inference of events happening close to the present), or we have a model of sequence evolution that can accurately account for the changes that have taken place over the evolutionary history of those sequences. Substitution rates inferred over long timescales using ancient viral sequences may therefore be an artificial construct influenced by two factors. First, which mutations are considered as fixed, and over which timescales. Second, (in the absence of evolutionary models able to account for all intermediate changes, or sequences representative of all changes during the evolutionary history of a virus), the timescale over which sequences are available for analysis, including the youngest and oldest sequences, as well as the distribution and diversity of sequences within that timeframe.

5.2 LIMITATIONS, FUTURE WORK, AND ETHICAL CONCERNS WHEN WORKING WITH ANCIENT VIRUSES

The results presented in the previous three chapters improve our understanding of the evolution of HBV, B19V, and VARV. Yet, there are a number of challenges when studying ancient viruses. In the following section, I expand on the limitations that studies on ancient viruses still face and avenues of future work.

First, as mentioned in chapter 1.3, the datasets that we screened for this thesis have originally been sequenced to study human population genetics. The viral sequences that are of interest to virologists, are only a by-product from the shotgun next generation sequencing that was used to generate the reads of the human genomes. Screening data from such studies is a powerful approach for virus discovery, since a large number of individuals have already been sequenced, and will continue to be sequenced in the future, thus increasing the chance of finding viral sequences. However, there are two major downsides of this approach when studying ancient viruses, which I have already touched on in chapter 1.2.1: one, current protocols for the sequencing of genetic material from ancient remains only allows for the sequencing of DNA, precluding us from finding ancient RNA viruses. Two, most likely only DNA viruses which cause deadly acute or chronic infections with high viral titres in the blood can be found in such data. This limits the viral species that can likely be detected using this approach to a subset of all human pathogenic viruses, that includes hepatitis B virus, human parvovirus B19, anelloviruses, orthopoxviruses, adenoviruses, and herpesviruses. In order to recover a more diverse set of viruses, different tissues could be sequenced, such as mummified tissue [113, 114], formalin-fixed, paraffin-embedded tissue [109], or other museum specimens [115]. Furthermore, it would be very interesting to attempt to sequence RNA viruses from ancient specimens, since RNA viruses exhibit rapid mutation rates, and can cause severe morbidity and mortality among humans [35, 138]. Experience in the laboratory suggests that RNA degrades more easily than DNA. The presence of cellular RNAses are assumed to quickly degrade RNA in a deceased individual. However, in conditions that disfavour RNase activity, such as aridity, RNA degradation is influenced by the chemical properties of the molecule [313]. The hydroxyl group at the 2' position of the ribose can attack the adjacent phosphodiester bond, resulting in hydrolytic cleavage of the RNA backbone. On the other hand, the hydroxyl group of the ribose increases the strength of the N-glycosidic bond that joins the ribose and the base, making RNA less susceptible to depurination and depyrimidation than DNA [313]. Finally, RNA is able to

form secondary structures, which may favour its preservation in some cases. Therefore, no clear evidence exists that RNA cannot persist over long periods of time under ideal preservation conditions [313]. The presence of ancient RNA sequences has been demonstrated in 1,400 year old cress seeds, as well as in an 723 year old maize kernel [314], and most recently, in permafrost-preserved liver tissue of a 14,300-year-old Pleistocene canid, and two wolf skins from before 1869 CE and 1925 CE [315]. Ancient viral RNA sequences have been sequenced by Smith *et al.*, 2014 [121], sequencing a 750 year old Barley Stripe Mosaic Virus, and by Peyambari *et al.*, 2018 [111], sequencing a 1000 year old plant virus. I am not aware of any human RNA virus sequences older than the A/H1N1 influenza virus from the 1918 pandemic that have been sequenced so far, and this is certainly be an exciting avenue for further research [316]. In addition to sequencing of ancient DNA and RNA, immunological (immunochromatographic detection, enzyme-linked immunosorbent assay, and immunohistochemical analysis) and analytical (mass spectrometry) methods could provide insight into the presence of viruses in ancient samples, especially in cases where nucleic acids cannot be recovered [57, 317].

Second, like studies based on ancient human genetics, studies of ancient viruses face limitations in terms of DNA preservation [63, 68]. Whether or not viral DNA is preserved better or worse than endogenous or environmental DNA is currently unknown. We also do not know whether ancient viral DNA generally persists inside a viral capsid, thus possibly better protecting it from degradation than the host DNA [118]. Krause-Kyora *et al.,* (2018) [112] identified HBV core proteins using liquid chromatography mass-spectrometry in samples ∼1000 and ∼7000 years old, suggesting that at least for HBV, core proteins may be stable in the long term.

Third, identification of viral sequences in metagenomic samples, ancient or not, relies on the presence of related sequences in public databases. Those databases are almost certainly an under-representation of the total viral diversity [20, 29, 30], and therefore it is possible that some ancient viruses are missed in the screening. Also, even though the evidence on HBV and B19V suggests otherwise, it may be possible that ancient viruses have diverged such that they cannot be detected using current sequence matching algorithms. The second part of this thesis introduces an unfinished approach to solve this problem, matching unknown sequences based on predicted secondary structures rather than by assign correspondence of nucleotides or amino acids between two sequences, as employed by most currently used tools.

Apart from the biological limitations, there are a number of ethical issues surrounding the work with ancient DNA and the viral sequences found therein.

First, ancient material is a limited resource and taking a sample from a mummy or skeleton for aDNA extraction is a destructive process. Apart from possible DNA, teeth and bones may contain important morphological information, which is lost from destructive sampling of ancient remains for DNA extraction. Care must thus be taken to make sure the maximum amount of knowledge can be gained from an ancient specimen, whether it is ultimately used for sequencing (be it through whole genome shotgun sequencing or targeted capture), for morphological or isotope analyses, or saved for later studies when technology may be more advanced.

Second, aDNA researchers have been criticised for being insensitive towards the communities and descendants of the individuals from which the samples are taken [318]. There is potential that the information gained from studying ancient genetic material negatively affects the identity of communities and individuals, disputes over territory and repatriation of remains, or leads to the release of stigmatising information, such as disease susceptibility [318]. Who (and if) should be asked for consent before sampling is often unclear, and may involve multiple stakeholders [318]. Since the sequences presented in this thesis are from individuals that are not directly culturally affiliated, this was not an issue here, but it may be in the future, should I begin to study viral sequences from individuals with direct ties to communities.

Finally, there may be concerns about biosecurity and biosafety around the work on ancient pathogens [34]. Some suggest that work on ancient samples may lead to the unintentional exposure to or accidental recovery and release of intact and infectious virus from ancient remains [34, 127]. Given what is known about decay kinetics of DNA over time [68], it would be highly surprising to find infectious virus in ancient remains. All viral sequences presented in the previous three chapters were assembled from reads that showed damage patterns typical for aDNA, suggesting that no live virus was present. Furthermore, no infectious VARV has ever been isolated, even though researchers have tried to do so repeatedly using a variety of sampling materials, including museum specimens, samples recovered from permafrost, smallpox scabs, and mummified material [319]. The same is true for unpublished attempts to isolate the A/H1N1 virus from the 1918 pandemic from four samples found in permafrost [320]. However, there has been one case of the recovery of a replication-competent Pithovirus *(Pithovirus sibericum)* from permafrost [127]. Given the quick degradation of DNA after the death of the host and the experience from the past decades, the risk of recovering a viable virus from ancient material which subsequently leads to infections seems to be negligible, but care should be taken when examining specimens where virus preservation may be suspected, such as samples found in permafrost [319, 321]. Furthermore, there may be concerns about the pos-

sibility that the work on ancient viruses could provide others with information that could be used to do harm, for example mutations leading to higher virulence, transmissibility, or vaccine escape. Such information could come either from the examination of the sequence directly (requiring some previous knowledge of what may be an interesting mutation), or from subsequent experiments performed on viruses resurrected based on an ancient sequence using reverse genetics. While researchers can learn some phenotypic information about a virus from its sequence (such as HBeAg status and serotype in HBV [242]), it is very difficult or impossible to draw conclusions about the virulence of a virus from the sequence alone, suggesting that the sequence information of ancient pathogens itself is of limited concern. In some cases, it may be possible to resurrect an ancient virus, or parts thereof, from the sequences that are recovered during aDNA studies. Tumpey *et al.,* (2005) have done so for the A/H1N1 virus from the 1918 pandemic [168]. In such cases, the potential risks from accidental or deliberate release of the resurrected virus and of possible findings about mechanisms for increased virulence or transmissibility, need to be weighted against the potential benefits, such as improved knowledge about the virus itself and its evolution, the continuing usability of vaccines, diagnostic tests and antivirals, and potential new targets for antiviral drug development. Phenotypic characterisation of ancient viruses must be evaluated carefully, taking into consideration the cost and benefits of doing and not doing the work [321].

5.3 CONCLUSION

The ancient viral sequences that were recovered from human remains provide a window into the past existence of three human pathogens. They reveal a surprising degree of sequence conservation over time, but also new and unexpected features. Further study of these ancient sequences will improve our understanding of virus biology, allow us to improve the models used to study virus evolution, and may lead to a better understanding of disease burden and mortality in the past. Future sequencing of ancient viruses from different sample types, animals other than humans, and of ancient RNA viruses, will further advance our understanding of virus evolution. The research area of studying viral sequences recovered in ancient remains is still in its infancy and future discoveries surely await.

PART II: AN ALGORITHM TO CLASSIFY SEQUENCES BASED ON PREDICTED STRUCTURAL FEATURES

CHAPTER 6: INTRODUCTION

The goal of sequence matching in the context of metagenomic analyses is to identify an unknown sequence by comparing it to one or multiple known reference sequences. This information can then be used for example, to identify the species in a sample (e.g., [86, 322, 323]), to estimate species abundance (e.g., [322, 324]), and to profile microbial community function (e.g., [324]). Sequence matching algorithms have been used to study the microbial composition of many environments, including patient samples [325], sea water [323], sewage [322], and ice cores [86], to name just a few. At least 70 tools for matching unknown sequences generated by NGS technology to reference genomes have been published since 2007, concurrent with the advent of NGS technology [26]. Methods for sequence matching can be divided into those that are alignment-based and those that are alignment-free [27,326]. Alignment-based methods compare sequences by looking for the correspondence between nucleotides or amino acids in the query[1] and the subject, with some leniency built in for mismatches and gaps, depending on the algorithm. Alignment-free methods do not produce an alignment. Instead, they compare the query and the subject by looking at measures not related to the correspondence of nucleotides or amino acids between the two sequences, such as the frequencies of subsequences of a predetermined length [326].

Methods for alignment-based sequence matching typically rely on either hash tables or the Burrows-Wheeler transform [26]. Algorithms based on hash tables first extract subsets of nucleotides or amino acids, called 'words', from the subject. A hash function converts each word to a location in the hash table, where the nucleotide or amino acid position of the word in the subject is stored. Words can either be contiguous stretches of nucleotides or amino acids of pre-determined length, or they can be 'spaced', where within a stretch of nucleotides or amino acids, only a subset is taken into account. In the second step, the algorithm extracts words from the query using the same method used for the subject, and screens for identical words in the query and the subject using the hash table. Subsequently, the matching algorithm extends the alignment from the initial matching words, using dynamic programming and an algorithm-dependent scoring function, to account for mismatches and gaps [26]. One of the most widely used tools for sequence matching, the basic local

[1] In the subsequent sections, I refer to the unknown sequence as the 'query' and the known reference sequence the query is compared to as the 'subject', following the language of BLAST [25].

alignment search tool (BLAST) and its variants[2], are based on hash tables [25, 26]. Some tools based on hash tables do not extend the alignments, and instead classify queries directly from the initial matching words. Examples are 'Kraken' and 'Taxonomer' [281, 327]. The absence of the extension step greatly increases the speed of these algorithms [281, 327]. Using simulated data with reads on average 92 bp in length, Kraken and Megablast classify 1.5 million and 7,143 reads per minute, respectively, with comparable accuracy [281]. Methods based on the Burrows-Wheeler transform include BWA [206] and Bowtie 2 [282]. The Burrows-Wheeler transform takes a subject and returns a sorted sequence and an index. The index allows fast look-up of queries in the subject [26]. Aligners based on Burrows-Wheeler transformations are generally faster than aligners based on hash tables followed by the extension of the alignment from the initial word match, since the costly extension step is mostly avoided [26].

Alignment-free methods quantify sequence similarity without taking into account the correspondence of nucleotides or amino acids in the query and the subject [326]. Most alignment-free methods are either based on the statistics of k-mer frequencies ('k-mer-based' methods), or methods that evaluate the information content between full-length sequences ('information theory-based' methods) [326]. Methods based on k-mer frequencies assume that similar sequences share similar k-mers (subsequences of length k). Such methods extract all possible k-mers of a predetermined length from the query and the subject, and construct k-mer frequency vectors for each sequence. The similarity between the two vectors can be measured using for example, the cosine of the angle between the two vectors. Alternatively, the distance between the endpoints of the vectors in high-dimensional space can be measured using a distance measure, such as Euclidean distance [326, 328]. Furthermore, k-mer frequency vectors of known sequences can be used to a train machine learning algorithm, such as a Naive Bayesian Classifier, to identify k-mers from an unknown sequence [329]. k-mer based methods include NBC [329] and TETRA [330]. Methods based on information theory investigate the amount of information shared between the query and the subject using measures of complexity and entropy. The complexity of a sequence can be measured using compression algorithms, as more complex sequences are less compressible [326]. If the compression of the concatenation of two sequences is similar to the compression of each sequence individually, the sequences are of

[2]Five different variants of BLAST exist (parentheses indicate whether the query (first element in the parentheses) or the subject (second element in the parentheses) are nucleotide or amino acid sequences): BLASTn (nucleotide – nucleotide, three varieties: Megablast, discontinuous Megablast, BLASTn), BLASTp (protein – protein, four varieties: BLASTp, PSI-BLAST, PHI-BLAST, DELTA-BLAST), BLASTx (nucleotide – protein), tBLASTx (protein – nucleotide, six frame translation), tBLASTn (nucleotide – nucleotide, six frame translation) [25].

similar complexity [326]. Entropy approaches measure the uncertainty of finding a given subsequence in a larger sequence (the rarer the subsequence, the higher its entropy) [326]. Alignment-free methods used in metagenomic analyses are typically based on *k*-mer frequencies [326].

Most studies attempting to identify viral sequences in NGS datasets use alignment-based tools. Nooij *et al.,* (2018) found that of 46 workflows published specifically for virus metagenomics, 43 used alignment-based tools (26 of those were BLAST variants) and three used profile Hidden Markov models (pHMM, see below). None used alignment-free methods [331].

While the screening of NGS datasets from many sources has uncovered a large diversity of previously unknown viruses [20, 29], there are still gaps in our knowledge of virus diversity, such as an almost complete absence of RNA viruses recovered from bacteriophages and amoeba [28, 29]. Our inability to find viruses in these parts of the tree of life may partly reflect the inadequacy of current virus discovery methods, alongside insufficient sampling and biological causes [28, 29]. When two homologous nucleotide or amino acid sequences have diverged such that the sequence similarity between them is close to the level that would be expected between random sequences, algorithms based on sequence similarity become formally unable to identify homologies without also producing an unacceptably large number of false positive matches, providing no hint as to how to differentiate between the false positives and the true positives. Sequence matching algorithms based on Hidden Markov Models and position-specific scoring matrices aim to address this problem of identifying sequences of homologous proteins that have diverged at the nucleotide or amino acid level. Tools include Position-Specific Iterative (PSI)-BLAST [332], and the HMMER [333] and HH-suite [334] software packages. PSI-BLAST uses the significant matches of the comparison of an amino acid query against an amino acid sequence database using BLASTp, to build a position-specific scoring matrix. The position-specific scoring matrix is then used to search for further matches, which are iteratively incorporated into the matrix. This allows the detection of similarity between amino acid sequences for which BLASTp is not able to identify statistical similarity, for example between the histidine triad proteins and galactose-1-phosphate uridylyltransferase proteins [332]. The HMMER and HH-suite packages both contain tools to build pHMMs from multiple sequence alignments, to compare them against reference pHMMs or reference sequences, and to compare queries against reference pHMMs [333, 334]. pHMMs are probabilistic models that represent the evolutionary changes in viruses based on a multiple sequence alignment. The pHMM contains states that represent sites in the multiple sequence alignment, with probabilistic state

transitions that summarise the insertions, deletions, mutations, and matches seen at the site. Databases of viral pHMMs, such as vFam and viral OGs, are available for the detection of viral sequences [335, 336]. pHMMs have been successfully used for the identification of contigs of viral origin and to detect remote homologies among protein sequences thought to be genus specific in RNA viruses [337,338]. However, they require long sequences to work well (sequences identified as viral using pHMMs by Bzhalava *et al.,* (2018) were on average 3,362 bp long) [337]. Furthermore, pHMMs rely on multiple sequence alignments which are based on nucleotide or amino acid sequences and require careful curation to adequately represent existing data and the evolutionary relationships therein [339]. pHMMs are therefore dependent on the currently available reference sequences, which most likely do not adequately represent the complete virus diversity [339]. Finally, some tools detect sequences that code for specific viral proteins. For example, ViralPro allows the detection of highly diverged phage proteins, such as the capsid and tail proteins, using a combination of pHMMs, amino acid and secondary structure composition [340]. The disadvantage of such tools is that they are of course restricted to a limited set of proteins and require long (>3000 base pairs) or complete protein sequences to work well [340, 341].

In summary, currently available methods for the screening of NGS datasets for viruses do not adequately address the problems posed by highly diverged viral sequences. While it is possible to identify sequences that correspond to a certain virus protein, like ViralPro does for the viral capsid and tail proteins [340], a general tool that can identify highly diverged viral sequences does not yet exist. To be of practical use, such a tool would need to be based on a computationally efficient algorithm, since NGS datasets and reference databases frequently contain millions or billions of sequences, and it would need to work well in the face of distortion due to mutations, insertions, deletions, and sequencing errors. Furthermore, the algorithm would ideally be able to deal with the short sequences (between 50 to 600 base pairs in length) generated by NGS technology.

6.1 TWO COMPONENTS OF A SEQUENCE MATCHING ALGORITHM

When developing an algorithm for sequence matching, we found it useful to think about two components separately: first, the alphabet that is used to compare sequences; and second, the algorithm that performs the comparison, based on the previously chosen alphabet. For example, BLASTn compares sequences based on a nucleotide alphabet, using a hash table and dynamic programming [25].

6.1.1 The alphabet

Current sequence matching algorithms typically use nucleotide or amino acid alphabets [26, 326, 331]. Alphabets can be based on features that correspond to biological function and that are easily interpretable by humans, but they need not be. What may look like a big difference using one alphabet may not be detectable using another. For example, a synonymous change will be visible using a nucleotide alphabet, but will not be visible using an amino acid alphabet. We can exploit this phenomenon for the comparison of highly divergent sequences. In order to detect relationships between sequences that have no or limited similarity at the nucleotide or amino acid level, one option is to think about the structure of the protein that the amino acid sequence folds into. The structure of a protein is more conserved than its sequence [32], and major secondary structure elements retain conformation, even under considerable sequence divergence [342]. Thus, proteins without sequence similarity at the nucleotide or amino acid level may still share a similar structure and it may be possible to exploit such structural information to detect highly diverged sequences. Two examples highlight the potential of incorporating structural information when studying viral sequences. By passing sequence as well as structural information to the phylogenetic tree inference algorithm MrBayes, the support of a phylogenetic tree of right-handed polymerases increases compared to a phylogenetic tree made from sequence or structural information alone [343, 344]. It has also been shown that using a stochastic model of sequence-structure evolution, phylogenetic trees can be calculated with significantly reduced alignment and topological uncertainty compared to a phylogenetic tree inferred using sequence information alone [345].

Using an alphabet that encodes information at the structural level may allow us to compare structurally homologous sequences with limited similarity at the amino acid level. To determine which structural features lend themselves to be incorporated into such an alphabet, we can consider if there are any discrete components that make up a protein structure. Protein structure refers to the three-dimensional arrangement of atoms in a protein molecule. The shape of a protein is determined by its amino acid sequence and by its environment. Four conceptual levels underlying protein structure are differentiated [346]:

Primary structure: the linear sequence of amino acids in a sequence.

Secondary structure: secondary structures are formed when local regions of the amino acid sequence fold into specific shapes. Protein secondary structures take the forms of alpha helices, beta strands, beta sheets, and loops. The secondary structure is determined by the pattern of hydrogen bonds formed between the amine hydrogen

and the carbonyl oxygen atoms of the backbone of an amino acid. Alpha helices are helical structures whose shape comes from the formation of regularly-spaced hydrogen bonds between the amine hydrogen and the carbonyl oxygen atom of the amino acid backbone. Three different types of alpha helices are distinguished, according to the number of amino acids per turn. The normal alpha helix has 3.6, the alpha helix 3-10 has three, and the alpha helix pi has five amino acids per turn. Helices are often positioned near the core of a protein. The part of the helix facing the inside of a protein usually consists of hydrophobic residues, while the amino acids facing outwards are hydrophilic. This leads to a typical repeated pattern of hydrophobic and hydrophilic amino acids in alpha helices [347]. Beta strands typically consist of three to ten amino acids. Beta strands have a propensity for valine, isoleucine, cysteine, phenylalanine, tyrosine, and threonine [348]. If hydrogen bonds are formed between the amine hydrogen and the carbonyl oxygen atoms of two different beta strands, they form a beta sheet. The strands in a beta sheet can be orientated in a parallel or anti-parallel arrangement. Loops do not have a defined structure and are typically found at the surface of the protein. Residue side chains of amino acids in loops generally do not form hydrogen bonds with other secondary structure elements.

Tertiary structure: the three-dimensional shape of a protein. The tertiary structure is produced by the folding of the secondary structure chain. It is governed by bonds between amino acid side chains, such as hydrogen, ionic, and disulfide bonds [346].

Quaternary structure: a protein with a quaternary structure is composed of two or more separate, folded amino acid chains, also called subunits, joined together by weak bonds. Not all proteins have a quaternary structure [346].

Apart from the shape of a protein, a number of other discrete structural features may be identified. These include amino acids such as cysteine, which are involved in the formation of disulfide bonds important for protein stability [349]. Further, regions important for the function of a protein, such as catalytic sites, and sites involved in the binding of metal ions, molecules or other proteins, are often short (3 – 20 amino acids in length) and highly conserved [349]. Thus, in order to detect a match (i.e., in our case, a structural correspondence) between highly diverged viral sequences, we can use an alphabet representing features that can be identified from an amino acid sequence, and that would be structurally conserved. These include the secondary structures described above, amino acids most likely to be conserved (tryptophan, cystein) [350, 351], sequence motifs [349, 352], and features based on amino acid properties.

6.1.2 The matching algorithm

Having decided on the features of the alphabet, we can think about the structure of the algorithm that could be used to perform the comparison. Before designing a sequence-matching algorithm to compare highly diverged biological sequences, it is worth considering whether similar problems have been solved in other fields. In particular, the problem of rapidly comparing a fragment of an unknown pattern against a large database of known patterns, and doing so in a way that is robust to differences that may be present between two related patterns.

6.1.2.1 An analogy with music matching

The field of Music Information Retrieval is concerned with retrieving information from audio signals. That information can then be used to compare pieces of music, classify songs by genre, identify the artist, automatically transcribe notes or lyrics, or search music collections [353]. Most relevant to the problem of biological sequence comparison are strategies developed in the field of Music Information Retrieval that enable the comparison of songs and song fragments. Broadly, two techniques can be used to compare and search audio signals: systems based on metadata, often using text-based search; and systems based on content, which compare audio signals based on features within the audio data [353].

Metadata-based systems: the most common way to access music is through textual metadata. For example, the internet radio station 'Pandora' relies on a manually curated database of songs and their attributes to suggest songs to listeners that they might like. Metadata can include the artist, song title, and genre, among others. However, as databases become larger, maintaining good metadata becomes an increasingly difficult task. For each song in Pandora's database, it takes an expert about 30 minutes to enter the metadata [353]. Some metadata based systems also have the drawback that they do not allow the user to search for music that they do not know how to construct a successful search for [353]. In the context of sequence matching, metadata-based systems are not relevant.

Content-based systems: content-based systems identify music based on the content of the audio. The algorithms generally have two processing stages: in the first stage, features are extracted from the audio recording, to reduce dimensionality and to achieve a simpler representation that is easier to compare. In the second stage, the features of two pieces of audio are compared, and distance between them is calculated [354]. Two types of features can be extracted from an audio recording,

generally described as high- and low-level features. High-level features use intuitive musical concepts such as melody, rhythm, and harmony to describe the content of the music [353]. Using high-level information allows, for example, query by humming applications, wherein music is identified and retrieved after the user hums a melody [353]. However, extracting these high-level features is often difficult, for example when multiple instruments are playing different melodies simultaneously. Low-level features are not necessarily meaningful in a perceptual or musical way and do not have to be interpretable by humans [354]. Such features can, for example, be spectrogram peaks, which are time-frequency points that have a higher energy than all their neighbours in a region centered around those points [355]. Others are spectral subband centroids or spectral flatness [354]. In either case, features need to be efficient to compute, robust to distortions and noise present in the audio, and must be able to identify the recording unambiguously [354]. After feature extraction, the feature vector of the unknown song or song fragment has to be compared against the feature vectors of known songs stored in a database. In order to quantify the difference between the two feature vectors, a distance has to be calculated based on a distance metric. Common distance metrics include correlation measures, Euclidean distance, and Hamming distance [354].

Content-based systems of music retrieval share many similarities with sequence matching. Firstly, it must be possible to compare a short audio fragment against a large database of known songs. This is similar to the problem of comparing short nucleotide or amino acid queries against a large database of known reference sequences. Secondly, the identification of an audio fragment must be possible even if there is loss of signal during recording or transmission, or ambient noise. Similarly, sequence comparison must work even if there is noise from mutations, sequencing errors, or recombination. Finally, given the high and increasing number of songs, and the large number of comparisons that are made, an algorithm for music matching must be fast. The same applies to sequence matching, where typically millions or billions of short NGS sequences are compared against reference sequence databases which are rapidly increasing in size. The similarities in the problems faced by content-based systems of music retrieval and sequence matching algorithms just mentioned, suggest that content-based music retrieval systems may offer a promising framework on which to model algorithms for sequence matching.

6.1.2.2 A closer look at the Shazam music-matching algorithm

Shazam is a commercial content-based music-matching algorithm, which implements matching based on audio features, and allows the identification of songs based on recordings of fragments of the original piece of music [355]. According to Avery Wang, the developer of the Shazam algorithm *'The algorithm is noise and distortion resistant, computationally efficient, and massively scalable, capable of quickly identifying a short segment of music captured through a cellphone microphone in the presence of foreground voices and other dominant noise, and through voice codec compression, out of a database of over a million tracks.'* [355].

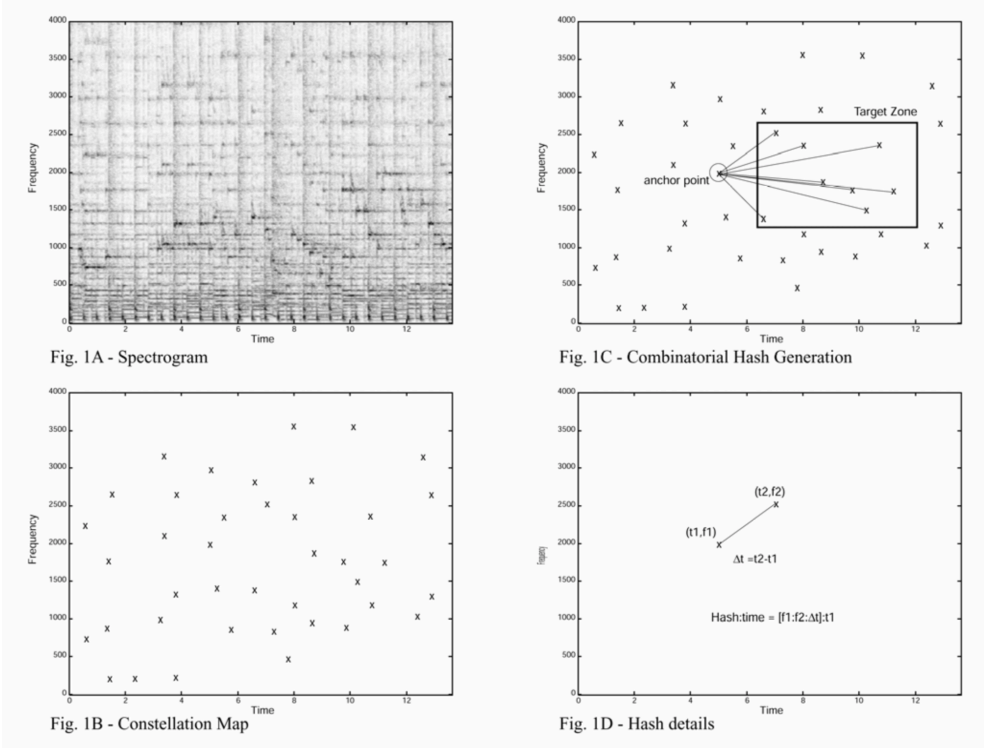


Figure 6.1: Feature extraction in the Shazam algorithm. Reproduced from Wang, 2003 [355].

According to Wang's publication from 2003 [355], the Shazam algorithm extracts features from each song in the database as well as from the incoming song fragment. The features are spectrogram peaks, which are peaks in a time-frequency plot (Fig. 6.1A). This reduces a complicated spectrogram to a sparse set of coordinates (Fig. 6.1B), which are easier to compare between songs. Shazam employs three criteria when selecting features: features should be close to each other in time, so that distant events do not affect the feature. They also need to be reproducible independent

of their position in the song, and robust to distortion due to recording or ambient noise. And finally, each feature must contain enough information to limit the probability of false positive matches. In order to increase the information content of each spectrogram peak, and to accelerate the search process, each detected peak is paired with another peak: to achieve this, 'anchor points' are chosen, and the anchor point is paired with points that are within a pre-defined target zone (Fig. 6.1C). Pairs of points and the time interval between them form an entry in a hash table and map to the time offset from the first of the two frequency points to the beginning of the song as well as the title of the song (Fig. 6.1D). The pairs of points in the song sample are compared to the pairs for each complete song in the hash table. For each matching pair, the time offsets into the song fragment and the complete song are subtracted and the resulting time differences are bucketed in a histogram. If there is a match between a song in the database and the song sample, there will be many matching pairs with an almost identical offset difference and the histogram will have a clear peak. Shazam can correctly identify songs in the presence of considerable levels of background noise. For a corrupted song fragment 15 seconds in length, the correct song can be identified even if only 1–2% of all pairs in the song fragment actually contribute to the match [355].

Shazam-like algorithms have recently been used for earthquake detection [356] and the analysis of sequence data generated by the Oxford Nanopore MinION sequencer [357], with reasonable and limited success, respectively.

6.1.2.3 Adaptation of principles of the Shazam algorithm to sequence matching

The adaptation of principles from the Shazam algorithm for matching biological sequences is attractive for many reasons. First, the problems are conceptually similar, with short fragments of linear patterns being compared against longer patterns. Second, the fact that the Shazam algorithm performs 20 million searches per day [358] is evidence that such algorithms can be scaled. Third, biological sequence comparison can be based on secondary structures or other features in the amino acid sequence that carry structural information. Identifying these with moderate accuracy may be relatively fast, and ensures that the structural features considered can be found within the short length of an NGS read. Fourth, due to the high number of entries in the database and the statistical nature of the matching, the identification of structural features does not have to be correct at all times. As with music matching, even a very high level of noise may not inhibit rapid and accurate matching. Finally, the envisaged algorithm

will not have to match features sequentially (in the order they occur in the sequence), and may therefore be more robust to insertions and deletions.

The factors outlined above allow for optimism that using a Shazam-like approach, it may be possible to construct an algorithm that will allow the matching of sequences based on predicted structural features. Chapter 7 of this thesis will provide an overview of the current version of the 'light matter' algorithm we have developed and chapter 8 presents preliminary results and discusses existing problems with the algorithm and avenues for its improvement. Details of methods for evaluating the algorithm during development, the structural features that were developed, methods for assessing the significance of a match, and methods for scoring a match, are in Appendix B. Additional results not presented in chapter 8 are in Appendix C.

CHAPTER 7: THE LIGHT MATTER ALGORITHM

This chapter describes the basic structure of the 'light matter' algorithm. It is modelled after the commercial music-matching algorithm 'Shazam', as published by Wang in 2003 [355], due to the similarity of the challenges in sequence and music matching that are outlined in the introduction of this part of the thesis.

The light matter algorithm takes amino acid sequences as input. In the following sections, I refer to the unknown sequence as the 'query', and the known sequence that the query is compared to as the 'subject', mirroring the terminology used by BLAST [25].

The basic steps of the light matter algorithm are:

1. Identification and pairing of features in query and subject sequences.
2. Matching of paired features between query and subject sequences.
3. Histogram binning of matching pairs based on relative offset differences.
4. Identification of significant bins.
5. Scoring of matches with significant bins.

These basic steps will be outlined in more detail below. For descriptions on the rationale and parameterisation of the algorithm, I include references to the relevant sections in Appendix B.

1) Identification and pairing of features in query and subject sequences (Fig. 7.1a).
In order to compare two sequences, we identify features in the sequences that we expect to be structurally conserved. Features must fulfil the following criteria:

1. Feature detection must be independent of the position of the feature in the sequence, to allow the matching of features in sequence fragments to those in complete sequences.
2. Features need to be robust to distortion from mutations, sequencing errors, or insertions and deletions.

3. Features need to be small enough so that it is possible to identify them in a sequence of around 65 amino acids, due to the short length of reads typical of Next Generation Sequencing (NGS) technologies.

In order to fulfil these criteria, we use secondary structures, sequence motifs, and amino acids that are conserved, as well as features based on amino acid properties. We divide features into two classes, landmarks and trigonometry (trig) points. Landmarks correspond to secondary structures (e.g., alpha helices and beta strands), sequence motifs, and conserved amino acids, whereas trig points are based on the sequence of amino acid properties and different conserved amino acids than those used by the landmarks. Once the features are identified, they are paired. We can enforce a minimum and maximum distance between features that are paired, to ensure that pairs do not involve features that are too close or too far from one another in the sequence. Furthermore, we can limit the number of pairs that each feature can be involved in. When pairing features, we allow a landmark to pair with a landmark or with a trig point, but we do not allow trig points to pair with each other. For each landmark, we first form pairs with all landmarks that are within the minimum and maximum distances specified, moving from closest to farthest, and we then form pairs with all trig points within the minimum and maximum distance specified, until the number of pairs allowed per landmark is reached. Twenty-two landmark and six trig point feature finders were implemented and are described in Appendix B.2.

2) Matching of paired features between query and subject sequences (Fig. 7.1b). For each pair in the query and the subject, the feature identities, the distance between the features, and the offset from the start of the sequence to the first feature in the pair are stored. The information for the subjects is pre-computed and stored in a database. The feature identities consist of a unique symbol for each type of feature, and optionally the length of the feature. Feature lengths and distances between features can be scaled logarithmically and the scaled lengths will then be used for further processing steps in the algorithm. This allows features to match, even if they do not have exactly the same length, and are not exactly the same distance apart. Feature lengths are by default scaled by using the integer part of the logarithm base 1.35 of the feature length and the distance between features is scaled by taking the integer part of the logarithm base 1.1 of the distance between features. The base of the logarithm can be varied by the featureLengthBase and the distanceBase parameters, respectively. Figure 7.2 shows the effect of scaling a distance of a particular length. Logarithmic scaling scales similar longer distances to the same distance, thus reducing the sensitivity of the algorithm. Finally, pairs in the query and the subject are compared. Two

pairs match, if they have the same feature identities, separated by the same (scaled) distance.

3) Histogram binning of matching pairs based on relative offsets (Fig. 7.1c, d).
If the subject and the query are similar, they should have multiple feature pairs in common, with similar distances between their relative offsets. We call the distance between their relative offsets the 'delta'. Deltas from all matching pairs are entered in a histogram. We can scale the deltas linearly (using the deltaScale parameter, Fig. 7.2) prior to entering them in the histogram, to allow matching pairs to fall into the same histogram bin, even if they do not have exactly the same delta. If the subject and the query have multiple feature pairs in common, at similar relative offsets (and hence with the same deltas), the histogram will have a peak, identifying a match.

4) Identification of significant bins.
To automate the inspection of histograms and assess the significance of a match, we need a quantitative metric to decide whether a histogram has one or more bins that are high enough to be indicative of a significant match. We calculate a significance cut-off defined as the theoretical maximum number of pairs that will be present in the highest bin if there is a perfect match between the subject and the query, multiplied by a user-defined significance fraction. Any bin that is higher than the significance cut-off is considered significant. Different methods for determining significance, as well as the rationale for choosing the current method to assess significance of bins, are outlined in appendix B.3.

5) Scoring of matches with significant bins.
After a match with at least one significant bin has been identified, the features in the pairs in the bin(s) need to be examined and assigned a score, reflecting the distance between the two sequences. Given that we want to compare sequences using predicted structural features, this score should correlate with the structural reality. Two types of scoring can be done, firstly by computing a score for each significant bin, and only reporting the highest score, a 'bin score', and secondly, by computing an 'overall score', which incorporates information from multiple bins. The scores are between 0.0 and 1.0, with 1.0 being the highest, representing full correspondence of feature pairs between the query and the subject.

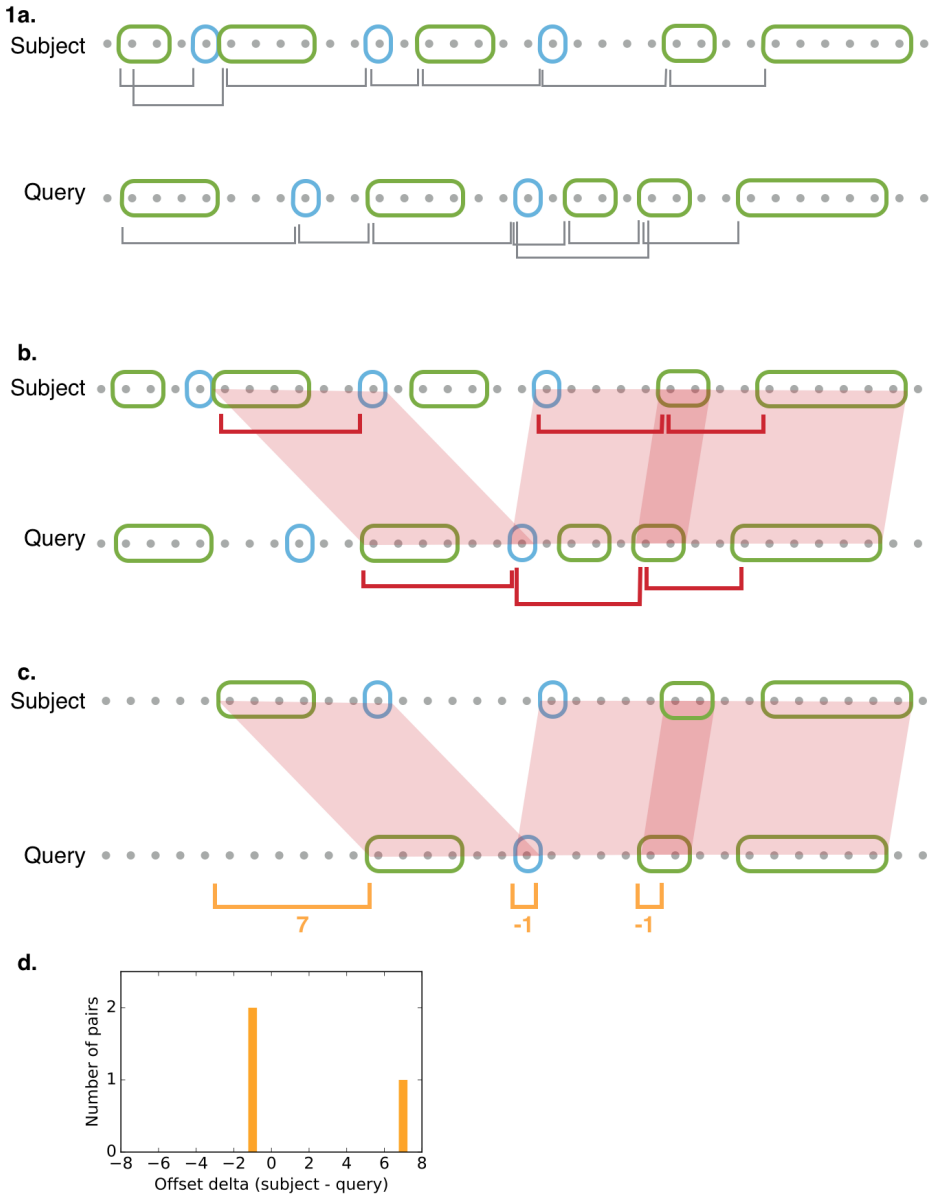


Figure 7.1: Step 1–3 of the light matter algorithm. a) Identification and pairing of features in query and subject sequences. Landmarks are shown in green, trig points in blue. Grey brackets indicate which features are paired, assuming a maxDistance of 7. **b) Matching of paired features between query and subject sequences.** Red brackets indicate matching pairs, the red shaded areas show which pairs match, darker red shaded areas indicate overlapping pairs. **c) Calculating relative offsets (deltas).** Deltas are calculated as the difference between the distances from the start of the sequence to the first feature of the subject and the query. In **c)** orange brackets indicate the size of the deltas, orange numbers their values. **d) The histogram of deltas resulting from matching the subject and query in this example.**

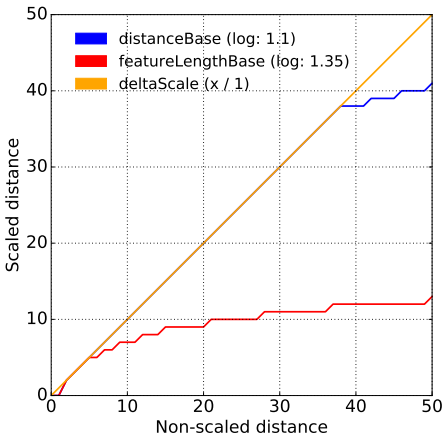


Figure 7.2: Scaling parameters. The distanceBase, shown in blue, scales the distance between two features in a pair using logarithm base 1.1 of the distance. The featureLengthBase (red) scales the length of a feature, where applicable, using a logarithm base 1.35 of the feature length. The deltaScale (orange) scales the deltas obtained by subtracting the offsets of two matching pairs, and here has been set to 1. Horizontally flat regions of the various lines indicate regions of equality (e.g., for histogram binning) to group similar raw values in order to reduce sensitivity due to feature locations or lengths that do not match exactly.

Three bin score methods were implemented (see Appendix B.4.1) and the one found to be the most appropriate, which we call the 'FeatureAAScore', is described here. The FeatureAAScore consists of a product of the score for the matched region and a length normaliser. The score for the matched region is a fraction where the numerator is the number of all amino acids in features in the subject and the query that match, and the denominator is the number of all amino acids in features in the subject and the query in the overall region spanned by all features in the bin. The length normaliser is the count of all amino acids in features in the region of the bin divided by the count of all amino acids in features in the whole sequence, for either the query or the subject, depending on which fraction is higher. Note that the bin with the highest score may not be the one with the highest number of pairs.

The overall score (called the 'GreedySignificantBinScore') is calculated using the same principle as the FeatureAAScore, except that multiple bins are taken into account. Bins are added to the overall score calculation sequentially. Significant bins are ordered by their FeatureAAScore. The overall score is first set to the FeatureAAScore of the best bin. Subsequent bins are then incorporated into the overall score until the overall score drops, at which point the overall score calculation concludes.

Technicalities

The light matter algorithm was written in Python 3.5, and was tested on Mac OSX (Yosemite, El Capitan, and High Sierra).

CHAPTER 8: RESULTS AND DISCUSSION

A lot of work has gone into the development of the light matter algorithm and its evaluation. In this chapter, I present what I consider the salient experiments that show the substantial difficulties in the approach of performing matching based on predicted structural features. To save the reader from the myriad details of the implementation of the feature finders, significance and scoring methods, as well as additional experiments on the evaluation of various aspects of the algorithm, I describe those in Appendices B and C. All of these experiments ultimately revealed substantial difficulties with the approach. Two years into the development of the light matter algorithm, I was given the opportunity to attempt to identify viral sequences in NGS datasets from ancient remains, which led to the work that now forms the first part of this thesis. On returning to the light matter algorithm when writing this thesis, I realise that different approaches for the identification of highly diverged sequences may now be more appropriate, given recent technical developments. These are mentioned in the final section of this chapter, together with a general outlook of the future of the light matter algorithm.

Evaluation of secondary structure finders based on test datasets of sequences with known structural similarity

Since the light matter algorithm performs sequence comparisons based on predicted structural features, it is obviously desirable that its scores correlate well with measures of structural similarity. To evaluate the algorithm from that perspective, I compiled five test datasets containing pairwise similarity measures of amino acid sequences based on both the structural similarity (measured by the Dali Z-score[1] [360]) and sequence similarity (measured by the BLAST bit score [361]).

The test datasets were constructed based on the sequence and structure information available from the Protein Data Bank (PDB)[2]. They consist of sequences of the major

[1] As a general rule, a Dali Z-score above 20 means the two structures are definitely homologous, between 8 and 20 means the two are probably homologous, between 2 and 8 is a grey area, and a Z-score below 2 is not significant [359].

[2] PDB is the primary repository for three-dimensional structural data on proteins and other biological macromolecules [362]. PDB stores the protein name, atomic coordinates of the structure, the sequence, the secondary structure element (if any) for each amino acid in the sequence, authors, key references, and derived data, such as quality assessment and fold classification [347].

histocompatibility complex (dataset '2HLA'), the RNA dependent RNA polymerase (datasets '4MTP' and 'Polymerase'), the virus capsid protein (dataset '4PH0'), and the influenza haemagglutinin protein (dataset 'HA'). For further details on the test datasets, see Appendix B.1.2.

The amino acid sequence and the correct secondary structure class (if any) of each amino acid can be acquired from all proteins in PDB. The secondary structure annotations can be used to evaluate the performance of the light matter algorithm when the secondary structures are identified correctly. This is achieved via the use of feature finders that use the secondary structure annotations in PDB to identify secondary structure features with perfect accuracy. I refer to these feature finders that are used in the following evaluations as the 'perfect finders' (Appendix B.1.1). The perfect finders make it possible to establish an important performance baseline for the algorithm, based on prior knowledge of structure taken from PDB. If secondary structure features are identified with 100% accuracy, how well does the algorithm do at finding overall structural similarity? If the results are only mediocre, there is a clear sign that the approach has some algorithmic limitations that are independent of the ability to identify secondary structures.

In addition to the perfect finders, three families of experimental finders that identify secondary structures are available to the light matter algorithm: the basic secondary structure finders identify alpha helices based on patterns of hydrophobic and hydrophilic amino acids in the sequence, and beta strands by screening for a sequence of at least six amino acids from the set [V, I, C, F, Y, T] [348] (Appendix B.2.1.1); the GOR4 finders identify alpha helices and beta strands based on the GOR4 secondary structure prediction algorithm (Appendix B.2.1.1); and the substring finders identify alpha helices and beta strands based on substrings of amino acid sequences that are known to occur frequently those secondary structures (Appendix B.2.1.1)[3].

I used the five test datasets in combination with the perfect finders and the three families of experimental finders to investigate the correlation of the scores assigned by the light matter algorithm ('light matter scores') with the Dali Z-scores. Figure 8.1 shows a grid of scatter plots, where the plots in each column show the light matter scores computed with different combinations of secondary structure finders plotted against the Dali Z-scores of that particular match. A linear regression was fitted to the data, with lines in green (for positive correlation) or red (for negative correlation), and its parameters given in the title of each subfigure. Rows of plots correspond to

[3]The names of the landmark finders that were used are: perfect finders: 'PDB AlphaHelix_combined' and 'PDB ExtendedStrand'; basic finders: 'AlphaHelix' and 'BetaStrand'; GOR4 finders: 'GOR4AlphaHelix' and 'GOR4BetaStrand'; substring finders: 'AC AlphaHelix_combined' and 'AC ExtendedStrand'.

different test datasets (top to bottom: 2HLA, 4MTP, HA, 4PH0, Polymerase). The first column shows the scores calculated when the secondary structures are identified correctly, using the perfect finders. The subsequent columns from left to right show the matches calculated with the basic finders, the GOR4 finders, and the substring finders.

In all but one of the finder combinations and test datasets in Fig. 8.1, there is a positive correlation between the light matter scores and the Dali Z-scores. The correlation between light matter scores and Dali Z-scores is best when the secondary structures are identified correctly (column one). Out of the three experimental finders, the substring finders (column four) lead to the best correlation between the scores computed by the light matter algorithm and the Dali Z-scores. Using the basic finders (column two) leads to a large number of matches with light matter scores of 0.0. The same is true to a lesser extent for the GOR4 finders shown in column three. The scores of 0.0 can be explained by the number of features that each set of finders identifies in the five test datasets: the basic finders identify a total of 275 features, the GOR4 finders identify 13,931 features, while the substring finders identify 56,550 features. This suggests that the basic secondary structure finders and the GOR4 finders are unable to identify enough features for a valid comparison.

While the correlation of the scores is best when the secondary structures are identified correctly, as shown in column one, there is also a wide range of light matter scores that are assigned to a narrow range of Z-scores and vice versa. For example, the second row in the first column of Fig. 8.1 shows that in the 4MTP test dataset, for comparisons with Z-scores in the region of 20, the light matter scores assigned range from about 0.3 to 0.6. Likewise, matches with a light matter score of 0.4 have Z-scores between 2 to 23. Also, looking at all datasets, there is little differentiation in the light matter scores assigned to comparisons with low Z-scores as opposed to high Z-scores. This is apparent when considering the scores in the HA test dataset, where all sequences have Z-scores of 20 and above. The lowest light matter scores assigned to matches in the HA dataset are approximately 0.4, while in the other datasets, light matter scores of that magnitude are assigned to matches with much lower Z-scores. The fact that these problems occur even with the correctly identified secondary structures, illustrates that the matching and scoring of the light matter algorithm, as currently implemented, must be improved to achieve acceptable results. All these problems are of course true to an even greater extent for the less accurate secondary structure finders.

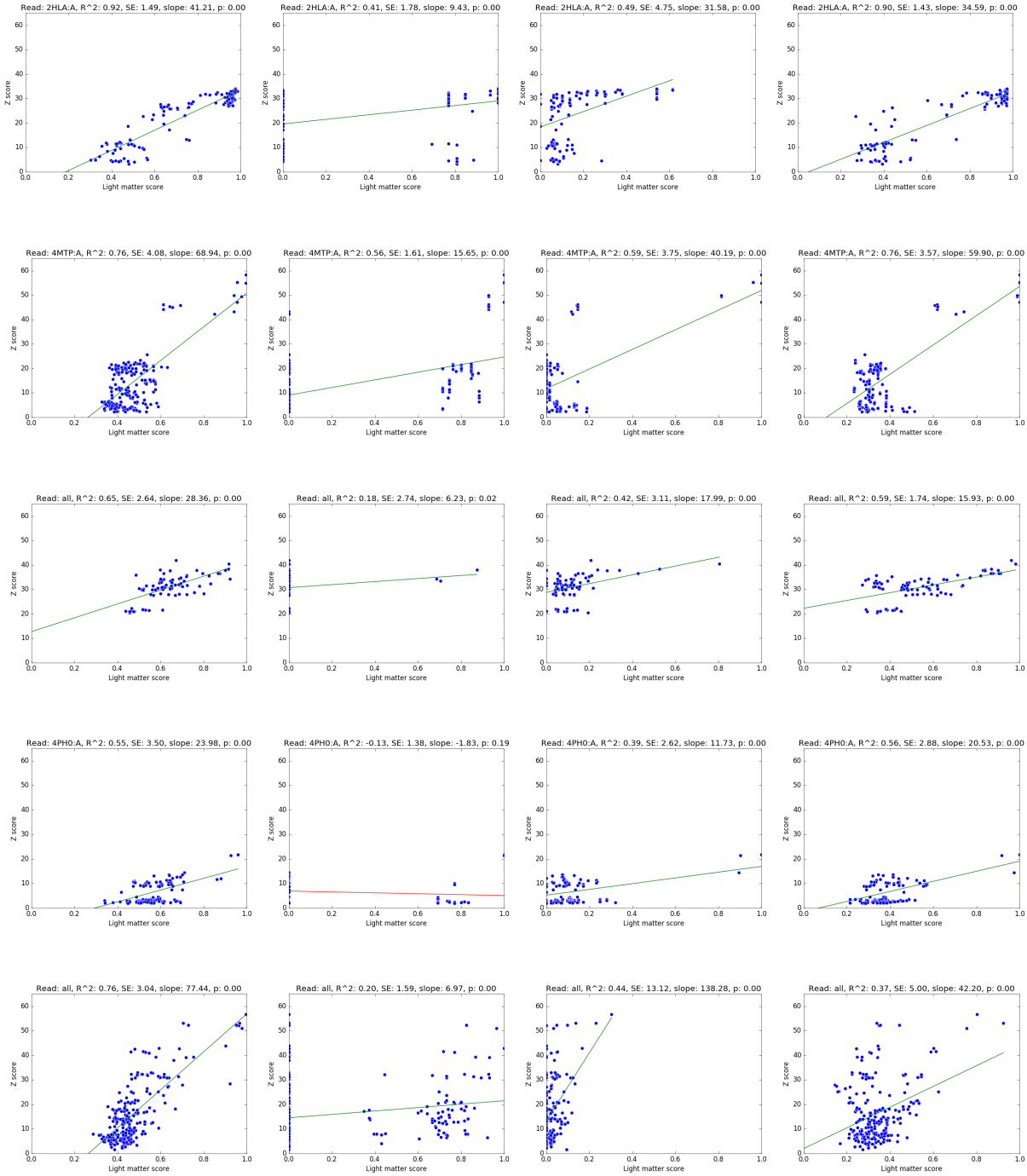

Read: 2HLA:A, R^2: 0.92, SE: 1.49, slope: 41.21, p: 0.00
Read: 2HLA:A, R^2: 0.41, SE: 1.78, slope: 9.43, p: 0.00
Read: 2HLA:A, R^2: 0.49, SE: 4.75, slope: 31.58, p: 0.00
Read: 2HLA:A, R^2: 0.90, SE: 1.43, slope: 34.59, p: 0.00
Read: 4MTP:A, R^2: 0.76, SE: 4.08, slope: 68.94, p: 0.00
Read: 4MTP:A, R^2: 0.56, SE: 1.61, slope: 15.65, p: 0.00
Read: 4MTP:A, R^2: 0.59, SE: 3.75, slope: 40.19, p: 0.00
Read: 4MTP:A, R^2: 0.76, SE: 3.57, slope: 59.90, p: 0.00
Read: all, R^2: 0.65, SE: 2.64, slope: 28.36, p: 0.00
Read: all, R^2: 0.18, SE: 2.74, slope: 6.23, p: 0.02
Read: all, R^2: 0.42, SE: 3.11, slope: 17.99, p: 0.00
Read: all, R^2: 0.59, SE: 1.74, slope: 15.93, p: 0.00
Read: 4PH0:A, R^2: 0.55, SE: 3.50, slope: 23.98, p: 0.00
Read: 4PH0:A, R^2: -0.13, SE: 1.38, slope: -1.83, p: 0.19
Read: 4PH0:A, R^2: 0.39, SE: 2.62, slope: 11.73, p: 0.00
Read: 4PH0:A, R^2: 0.56, SE: 2.88, slope: 20.53, p: 0.00
Read: all, R^2: 0.76, SE: 3.04, slope: 77.44, p: 0.00
Read: all, R^2: 0.20, SE: 1.59, slope: 6.97, p: 0.00
Read: all, R^2: 0.44, SE: 13.12, slope: 138.28, p: 0.00
Read: all, R^2: 0.37, SE: 5.00, slope: 42.20, p: 0.00
Z score
Light matter score

Figure 8.1: Correlation of light matter scores and Dali Z-scores, using different combinations of landmark finders. Each row corresponds to a different test dataset, top to bottom: 1) 2HLA. 2) 4MTP. 3) HA. 4) 4PH0. 5) Polymerase. Each column corresponds to a different combination of landmark finders, from left to right: 1) Perfect finders. 2) Basic finders. 3) GOR4 finders. 4) Substring finders. The 'Peak', 'Trough', and 'AminoAcids' trig point finders, a significance fraction of 0.01, a FeatureLengthBase of 1.01, DistanceBase of 1.1, DeltaScale of 1.0, and a MaxDistance of 1000 were used throughout. Lines indicate the linear regression, with coefficients in the title of each subfigure. A green regression line indicates a positive correlation, a red regression line a negative correlation.

Evaluation of secondary structure finders based on visualisations of the three-dimensional protein structure

While the examination of correlations between the light matter and Dali Z-scores provides a broad overview of the performance of the light matter algorithm, it tells us nothing about the factors involved in the calculation of the light matter score. It is therefore necessary to investigate matches of interest in isolation, for example by using visualisations that show the location of features on the three-dimensional protein structures and how the features match between structures according to the light matter algorithm. This complements the investigation above: not only do I want to know if features are being identified, but also whether the matching of the identified features between sequences corresponds to reality. This may help us to understand the discrepancies between the light matter scores and the Dali Z-scores observed in Fig. 8.1. In the following section, I will focus on the discrepancy in four example matches. Two have low light matter scores but high Z-scores, and two have high light matter scores and low Z-scores. For each match, I show the two structures next to each other, with the features identified by the substring finders highlighted[4].

Figures 8.2 and 8.3 show two examples for the comparison of structures that have Z-scores that indicate that the sequences are definitely homologous (>20) [359], but that are assigned low light matter scores (0.21 and 0.3). Examining the three-dimensional structures in Figs. 8.2 and 8.3, shows that they look similar visually, with a similar arrangement of secondary structures, as indicated by the high Z-score assigned by the Dali algorithm. The figures also highlight the features identified by the substring finders. Features identified as alpha helices are in brown, features identified as beta strands in green, and features that match between the two structures in red. The colouring shows that many secondary structure elements are not identified as features at all, or that alpha helices are identified as beta strands and vice versa. Furthermore, the features that are matching between the two structures, as indicated in red and by white connecting lines, often do not belong to secondary structures that are structurally equivalent (i.e., for example, a pair containing an alpha helix matches another pair containing an alpha helix, but the two alpha helices are from different regions in the two structures). The light matter scores in this case does not accurately reflect the structural similarity, as it is based on an incorrect correspondence between secondary structures.

[4]In addition to the landmark finders ('AC AlphaHelix_combined' and 'AC ExtendedStrand'), the following parameters were used: a significance fraction of 0.01, a FeatureLengthBase of 1.01, DistanceBase of 1.1, DeltaScale of 1.0, and a MaxDistance of 1000.

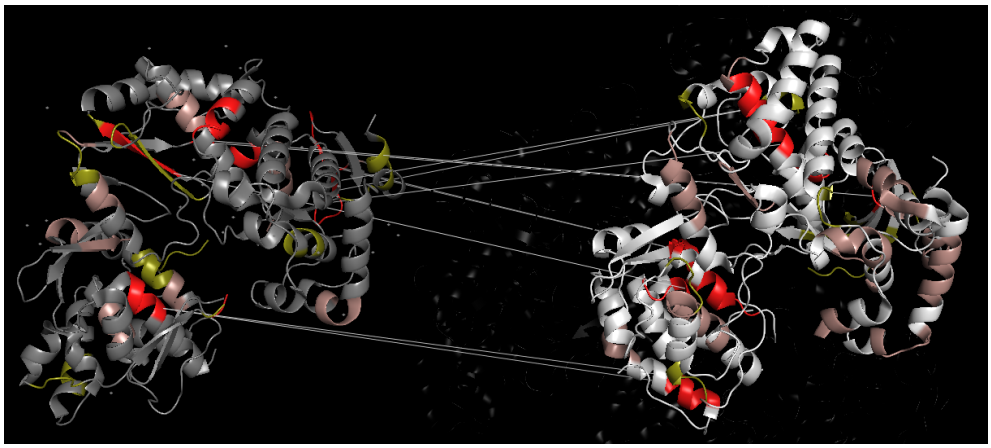


Figure 8.2: Structures of 4MTP (right) and 2CJQ (left). 'AC AlphaHelix_combined' in brown, 'AC ExtendedStrand' in green, matching features in red. White lines show features in matching pairs. FeatureAAScore light matter score: 0.3. Dali Z-score: 25.6.

In Figs. 8.4 and 8.5, I show two examples of matches with relatively high light matter scores of 0.47 and 0.61, and Z-scores of 3.0 and 3.5. Why does the light matter algorithm assign scores that are seemingly too high? In the two matches shown, the sequences coding for these protein structures are of different lengths (634 versus 308 amino acids for 4MTP and 2G1H, and 634 versus 98 amino acids for 4MTP and 1H6K), and the structures are not visually similar. The visual impression thus agrees with the low Z-scores. The length discrepancy between the sequences that code for the structures can artificially increase the light matter score. The length normalisation that is performed in the FeatureAAScore calculation rewards matches where the matching pairs are distributed across the entire query and / or subject. Indeed, for the comparisons in Figs. 8.4 and 8.5, the length normalisers are 0.898 and 0.821, respectively. Furthermore, in both cases, there are relatively few non-matching features within the matched region, which also leads to high matched region scores (0.512 and 0.879, respectively). Finally, as in the examples in the previous section, the location of the matching features (as indicated by white lines in Figs. 8.4 and 8.5) shows that they are often not structurally equivalent.

Limitations of the light matter algorithm and its evaluation

The evaluations based on the correlation between light matter scores and Dali Z-scores, as well as the more detailed investigation of a small number of matches using visualisations of the three-dimensional structures, show that the performance of the light matter algorithm using features based on secondary structures is not good

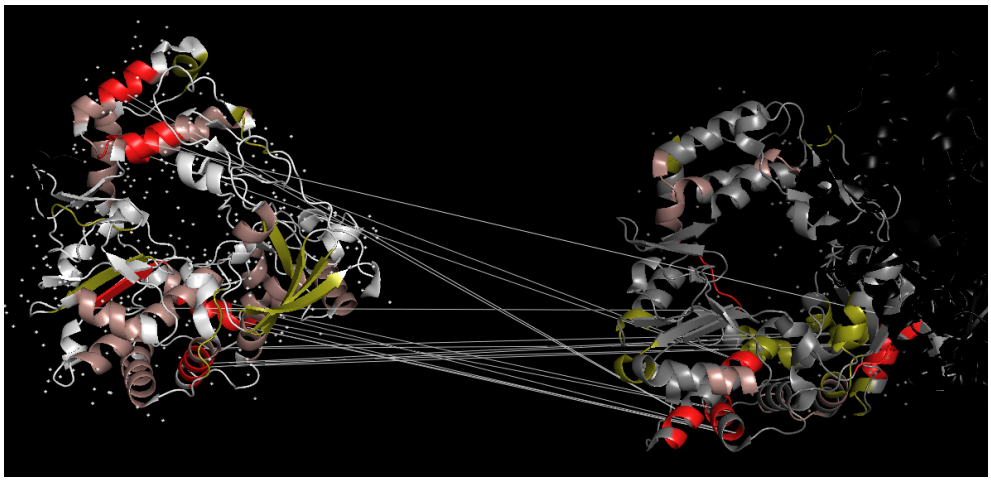


Figure 8.3: Structures of 3CDW (right) and 1KHV (left). 'AC AlphaHelix_combined' in brown, 'AC ExtendedStrand' in green, matching features in red. White line show features in matching pairs. FeatureAAScore light matter score: 0.21. Dali Z-score: 33.0.

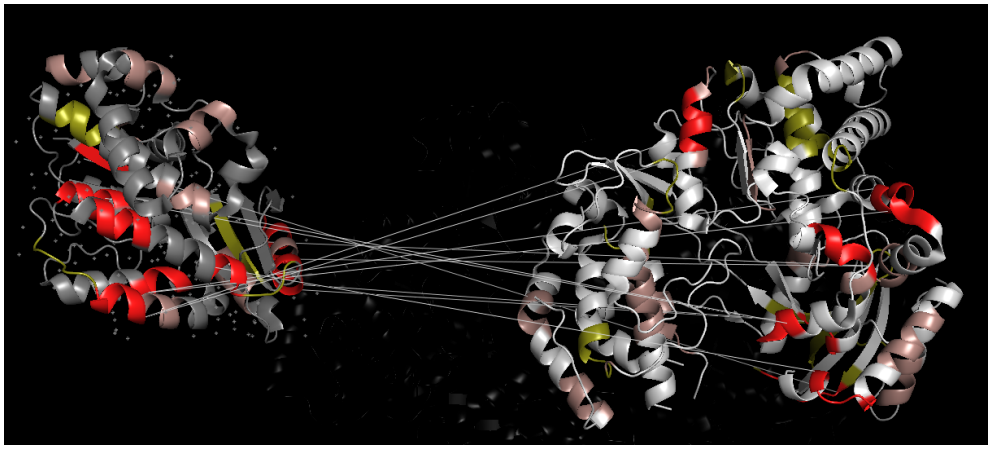


Figure 8.4: Structures of 4MTP (right) and 2G1H (left). 'AC AlphaHelix_combined' in brown, 'AC ExtendedStrand' in green, matching features in red. White lines show features in matching pairs. FeatureAAScore light matter score: 0.47. Dali Z-score: 3.0.

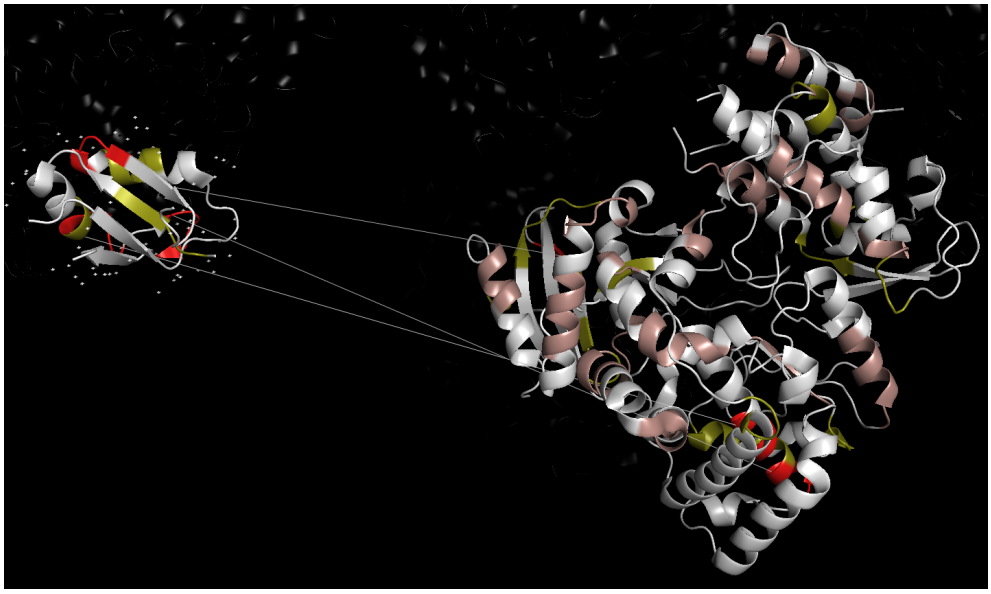


Figure 8.5: Structures of 4MTP (right) and 1H6K (right). 'AC AlphaHelix_combined' in brown, 'AC ExtendedStrand' in green, 'AminoAcidsLm' in pink, matching features in red. White lines show features in matching pairs. FeatureAAScore light matter score: 0.61. Z-score: 3.5.

enough for its intended purpose. The evaluations highlight that the light matter algorithm has limitations due to difficulties with secondary structure identification and matching of corresponding features between structures. Furthermore, there may also be limitations to the approach of evaluating the light matter algorithm, in particular due to the reliance on measures of structural similarity based on tertiary structures, as inferred by the Dali algorithm.

Limitations of the algorithm due to secondary structure identification

The evaluations presented earlier show that the performance of the light matter algorithm using features based on secondary structures is not adequate, for the following reasons:

– The basic secondary structure finders and finders based on the GOR4 algorithm do not identify a sufficient number of features based on secondary structures to allow for matches to be identified, which results in light matter scores of 0.0, even when there is close structural similarity. Furthermore, the visual examination of matches with lower than expected light matter scores (Figs. 8.2 and 8.3) suggests that the substring finders also do not always identify sufficient secondary structures.

– The visual inspection of matches with light matter scores lower or higher than expected shows that even if secondary structure features can be identified, there may be matches between secondary structures that are not structurally equivalent. This can lead to matches with lower than expected scores on the one hand, and may also lead to matches with scores higher than expected on the other.

– Finally, while not directly related to the identification of secondary structures, the score calculation may be flawed in cases where a short sequence is compared to a long one (an artefact of the FeatureAAScore calculation, Appendix B.4).

The broad correlations between the light matter scores and the Dali Z-scores suggest that the algorithm is able to identify structural relatedness. But the frequent occurrence of light matter scores lower or higher than expected show that the performance of the light matter algorithm is not currently good enough for its intended purpose.

Limitations of evaluation due to reliance on Z-scores

When evaluating the light matter algorithm, I relied heavily on the Z-scores computed by the Dali algorithm as a measure of structural similarity. However, structural similarity measured by comparing tertiary structures, as the Dali algorithm does, may not always accurately reflect how similar two structures are. Apart from quantitatively measuring the distance between tertiary structures, structural similarity is often discussed subjectively. Protein structures that are subjectively considered to be similar may not be similar when the distance between them is measured quantitatively. This may be the case when two structures are compared and one of them has a ligand bound to it (e.g. Fig. C.8 in Appendix C.2), or is in a different conformation. A further example is the proposed similarity between the Herpes Simplex Virus Glycoprotein B and the Vesicular Stomatitis Virus G protein (Fig. 8.6), which are subjectively described as *'so similar'* and *'Glycoproteins from two entirely different viruses share the same novel structure,...'* [363]. Indeed, they share the same domains with a similar arrangement of secondary structures. However, the Dali algorithm assigns a Z-score between 2.4 and 3.7, depending on the chains that are compared, and BLASTp reports a bit score of only 19.6 between chains of the two structures. Thus, according to the two quantitative measures, there is no structural or amino acid sequence similarity between the two proteins, in contrast to the subjective assessment.

Given that the light matter algorithm performs comparisons based on predicted secondary structure features, there could be scenarios where it detects similarity between

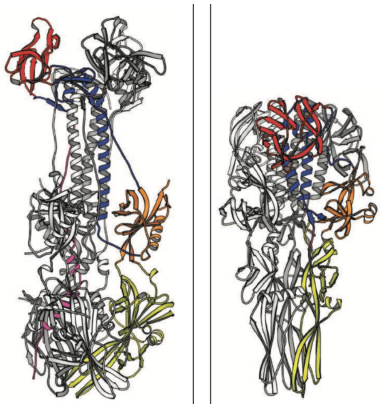


Figure 8.6: Subjective structural similarity Ribbon diagrams of the ectodomain trimers of glycoprotein B of Herpes Simplex Virus (left) and G protein of Vesicular Stomatitis Virus (right). On a single subunit, corresponding domains are coded in yellow, orange, blue, and red; the extra domain V on glycoprotein B is pink. In both cases, the other two subunits are white and pale grey (Figure and caption modified from [363]).

two structures that is not detected by the Dali algorithm. Should we ignore such a match on the basis of a low Z-score? Or could the light matter algorithm be a tool to detect similarity between sequences coding for structures such as the ones in Fig. 8.6, that cannot be detected by the Dali algorithm (should it perform well enough in the future)? I would argue in favour of the similarity detected by the light matter algorithm, since the algorithm is not sensitive to changes in the conformation of the tertiary structure. However, systematically evaluating this behaviour would require a way of evaluating the light matter algorithm that does not rely on structural similarity measured at the tertiary level. For example, visually inspecting the three-dimensional structures of the sequences that are being compared would be helpful to get a more detailed insight into the equivalence of two structures, which may not be reflected in the Dali Z-score, but would also be impractical to do at a large scale.

Future work

The evaluations presented in this chapter show that the features on which the sequence comparison is based in the light matter algorithm needs to be improved. Currently, features are based on predicted secondary structures and other structural features. It may be possible to make improvements at this level, by implementing features based on profile hidden Markov models of complete viral proteins, as implemented in the vFam database [335], or partial sequences of viral proteins [339]. In addition, features that are less descriptive of the protein structure could be implemented. For example, the amino acid sequence could be transformed into a numerical vector, using representations such as the electron-ion potential, or the Voss representation [364]. This vector could then be analysed using time series data mining techniques such as the discrete wavelet transform [364], to provide features for

the light matter algorithm. The light matter algorithm is built in a modular fashion, making it easy to add additional feature finders, thus facilitating future collaborative development. The examination of the three-dimensional structures of matching sequences showed that there are frequent matches between pairs of features that are not structurally equivalent. It may be possible to alleviate this problem by looking for identical triplets of features between two sequences, instead of pairs of features. This would make matches more specific, but false negative features would also have a bigger impact. Matching with triplets would also lead to new challenges with the implementation of triplet formation, and would probably lead to a bigger database size due to the increased number of possible feature combinations.

Outlook for the light matter algorithm

While the current results presented in this chapter do not exclude that it will eventually be possible to implement adequate feature detection and scoring for the light matter algorithm, it has not been possible to do so up to now. Therefore, the obvious question to ask at the end of this chapter is whether there is any indication that the algorithm will eventually be useable for detecting highly diverged, homologous viral sequences. The success of the 'Shazam' application shows that the overall approach of the algorithm is effective and highly scaleable [355], which is encouraging. Whether it is possible to develop adequate features to base the sequence comparison on, is more difficult to predict, and is most likely the area in which improvements will have the strongest effect on the performance of the algorithm.

One approach to make the light matter algorithm useful in the short term, is to attempt to initially make it work well on just a selected class of viral proteins, instead of attempting to implement an algorithm that is of completely general applicability. Features could be selected and the algorithm parameterised to identify a specific group of proteins, such as the Jelly-roll capsid protein, the Superfamily 3 helicase, or the RNA-dependent RNA polymerase [365]. Such an approach would probably be the most realistic and quickest way for the algorithm to become useable, but it is also a problem that has already been solved by others (e.g., [340]). Another option may be to apply the algorithm to problems other than the identification of unknown sequences. Features detected by the algorithm could be used as additional information for the classification of known sequences, or could be incorporated into phylogenetic analysis to possibly increase support in trees. Therefore, while this project was difficult and ultimately without clear results and success, some level of optimism remains.

Work on the light matter algorithm started in late 2014. Given that more than four years have passed since then, another question is whether the approach of matching sequences based on predicted structural features is still a promising approach for the identification of sequences of unknown origin and therefore worthy of additional work. This consideration is especially important in light of improvements in sequencing technology and protein structure prediction during the last four years. Single molecule sequencing technology, commercialised by Oxford Nanopore Technologies and Pacific Biosciences, is now widely available and allows the routine production of much longer sequencing reads, making it possible to directly sequence viral nucleic acids without an amplification step [366]. Long reads or full genomes would improve the performance of the light matter algorithm, but such sequences could also be classified more easily using other methods, such as profile hidden Markov models or by exploiting information about gene arrangement [339, 340]. Furthermore, at the end of 2018, it was announced that the AlphaFold software won the 13$^{\text{th}}$ Critical Assessment of Structure Prediction, a bi-annual competition on protein structure prediction. The advance in prediction accuracy achieved by AlphaFold was an advance of roughly double the improvement achieved in the previous competitions in 2014 and 2016 [367, 368]. Further progress in protein structure prediction, which may come in the near future, combined with longer sequences available from single molecule sequencing technology, might turn the problem of sequence comparison into a problem of comparing structures, which will require different approaches to performing the comparison and interpreting the results than those which the light matter algorithm uses. Further improvements in sequencing technology and structure prediction will therefore most likely improve the results of the light matter algorithm, but may also render it obsolete in the future.

CONCLUSION

This thesis highlights both the benefits and the outstanding issues when screening next generation sequencing (NGS) datasets for viruses.

The first part of the thesis showed the potential of viral sequences recovered from screening NGS datasets from individuals living as far back as the late Pleistocene, to improve our understanding of virus evolution. The ancient hepatitis B virus, human parvovirus B19, and variola virus sequences reveal some of the diversity of those viruses over the past hundreds or thousands of years, and also provide evidence for genotype extinction. They allowed us to re-estimate most recent common ancestor dates and substitution rates, and in some cases make inferences about the possible correlations with human migrations. Furthermore, the lower substitution rates we inferred with the ancient sequences, as compared to just using modern sequences, highlight some of the problems with the current thinking about substitution rates. Studying virus evolution with sequences recovered using techniques developed in the field of aDNA research is an area of research still in its infancy. The work presented in chapters 2 – 4 of this thesis represents some of the first large-scale studies on ancient viral sequences, and is mainly descriptive of the sequences recovered in the datasets. In the future, in addition to screening of datasets that will be generated for studies on human population genetics, targeting specific historical or virological questions will become important. For instance, elucidating the origin and spread of HBV genotypes F and H in the Americas, or the origin of the modern non-human primate HBVs are possible questions that we may be able to answer. In addition, aDNA techniques could be applied to identify putative causative agents of past epidemics described in the historical literature, such as the Antonine plague. Furthermore, sequencing of ancient RNA viruses, if possible, will most certainly lead to exciting new insights.

The ancient sequences of hepatitis B virus, human parvovirus B19, and variola virus presented in the first part of this thesis, were identified using standard sequence matching tools. Such tools mainly compare sequences at the nucleotide and amino acid level, and are unable to detect similarity between sequences that may code for

proteins with similar structures, but without apparent nucleotide or amino acid sequence similarity. In the second part of the thesis, I describe an attempt to develop an algorithm to match sequences based on predicted structural features, with the goal of being able to identify highly diverged sequences that still retain similarity at the structural level. The development of the algorithm was ultimately not successful, in part due to difficulties in feature identification. Advances in single-molecule sequencing technology and structure prediction since the start of the development of the algorithm in late 2014, may allow researchers to develop different and possibly more effective approaches for identifying highly diverged sequences in the future.

PART III: APPENDICES

APPENDIX A: ADDITIONAL TEXT TO 'DIVERSE VARIOLA VIRUS LINEAGES CIRCULATED IN NORTHERN EUROPE DURING THE VIKING AGE': DESCRIPTION OF GENE FUNCTIONS

In the following text, numbers correspond to the final offset of the gene in CPXV-Gri/GER [270] and the name of the VACV-COP homolog [270], if applicable. If the gene is absent in VACV-COP, the CPXV naming convention is used.

Gene status is organized into eight categories, denoted A–H.

VARV and VARV-VD21 gene status	Category: aVARV gene status (gene count)
Inactivated	A: inactivated in all (11)
	B: certain or uncertain inactivations in some but not all (5)
	C: present in all (4)
	D: present in some but not all (9)
	E: uncertain inactivations (3)
Present	F: absent in some (2)
	G: uncertain inactivations in some (17)

Table A.1: Gene inactivation and presence categories. Gene status is divided into eight categories A–H. The table shows categories A–G and the number of genes in each. For clarity, categories A–G are organized in two groups, according to whether the gene is inactivated or present in VARV and VARV-VD21. An eighth category, H (with gene count 3), is used for genes that are absent or with no coverage in VARV, VARV-VD21, and aVARV sequences, and that have gene-inactivating mutations in CMLV or TATV.

A) Genes inactivated in VARV, VARV-VD21, and the aVARV sequences.

This includes 11 genes, *A25L* (150930), *A39R* (165419), *A37R* (163060), *C8L* (31175), 20088, *B2R;B3R* (182986), 26765, *A9L* (135109), 202711, *C13L;C14L* (204801), and *A53R* (176775).

A25L: A-type inclusion protein [270]. VARV-VD21 and all aVARV sequences share their gene-inactivating mutations with modern VARV and CMLV. Coverage of reference sequence (TATV): VARV-VD21: 100%, aVARV-VK382: 94%, aVARV-VK388: 99.8%, aVARV-VK470: 99.2%.

A39R: Semaphorin, secreted glycoprotein. Immunomodulator [7, 270]. Loss of the gene from VACV-COP does not affect growth in vitro, or virulence in a mouse intranasal model, and has only a slight effect on lesion size in an intradermal mouse

model. Expression of VACV-COP *A39* by VACV-WR, which does not naturally express the gene, leads to an increase in the severity and persistence of skin lesions after intradermal infection of mice [369]. Histological examination of mouse skin suggests that *A39* has direct or indirect pro-inflammatory properties [369]. aVARV-VK388 and VACV-WR share the same gene-inactivating mutation. aVARV-VK382 has a 1 nucleotide (nt) deletion, resulting in a novel stop codon. aVARV-VK470 also has a novel stop codon. Coverage of reference sequence (TATV): VARV-VD21: 100%, aVARV-VK382: 60.9%, aVARV-VK388: 100%, aVARV-VK470: 100%.

A37R: Unknown [270]. Modern VARV, VARV-VD21, and all aVARV sequences share the same stop codon. Coverage of reference sequence (CPXV-BR): VARV-VD21: 100%, aVARV-VK382: 91.5%, aVARV-VK388: 100%, aVARV-VK470: 24.1%.

C8L: Unknown [270]. A novel, identical stop codon in VARV-VD21, and all aVARV sequences. Coverage of reference sequence (CPXV-BR): VARV-VD21: 100%, aVARV-VK382: 91.2%, aVARV-VK388: 100%, aVARV-VK470: 95.5%.

20088: CPXV019. Ankyrin [270]. The aVARV sequences share a stop codon caused by a 1 nt insertion. There are additional stop codons in the sequence, one of which agrees with the stop codon in VARV. Coverage of reference sequence (TATV): VARV-VD21: 14.3%, aVARV-VK382: 77.4%, aVARV-VK388: 100%, aVARV-VK470: 100%.

B2R;B3R: Schlafen. Immunomodulator [270]. aVARV-VK382, aVARV-VK388, and aVARV-VK470 share the same stop codon. Coverage of reference sequence (CMLV): VARV-VD21: 45.2%, aVARV-VK382: 41.7%, aVARV-VK388: 45.1%, aVARV-VK470: 45.8%.

26765: CPXV025. ANK/F-box protein related to host range [273]. Identical stop codons in all aVARV sequences. Coverage of reference sequence (VARV-SAF): VARV-VD21: 100%, aVARV-VK382: 97.8%, aVARV-VK388: 100%, aVARV-VK470: 99.6%.

A9L: Membrane protein [270]. Differing stop codons in all aVARV sequences. VARV-VD21 agrees with VARV-BRZ [273]. Coverage of reference sequence (VARV-BRZ): VARV-VD21: 96.2%, aVARV-VK382: 100%, aVARV-VK388: 100%, aVARV-VK470: 100%.

202711: CPXV215. Kelch-like protein [270]. VARV-VD21, aVARV-VK382, and aVARV-VK470 share a gene-inactivating mutation with VARV. aVARV-VK388 does not have coverage in later part of the sequence, and may agree with VARV too. Cov-

erage of reference sequence (CPXV-Ger): VARV-VD21: 49.4%, aVARV-VK382: 31.8%, aVARV-VK388: 48%, aVARV-VK470: 49.2%.

C13L;C14L: Unknown [270]. VARV-VD21, aVARV-VK382, aVARV-VK470, and VARV share the same gene-inactivating mutation. aVARV-VK382 and aVARV-VK470 also have novel stop codons toward the N-terminus of the sequence when aligned against TATV. aVARV-VK388 has a novel stop codon mutation. Coverage of reference sequence (TATV): VARV-VD21: 100%, aVARV-VK382: 100%, aVARV-VK388: 100%, aVARV-VK470: 100%.

A53R: Also called *CrmC*, is a tumor necrosis factor receptor homolog that plays an important role in antiviral response and inflammation [273]. Expression of the CPXV or VACV-USSR A53R gene in VACV-WR leads to increased virulence in a mouse intranasal model [370]. Deletion of *A53R* in Tiantian VACV does not lead to a change of lesion size upon intradermal infection of rabbits, but resulted in virus attenuation in the mouse intranasal and intracranial model [371]. Gene-inactivating mutations agree in all aVARV sequences. Coverage of reference sequence (CPXV-Ger): aVARV-VK382: 80.6%, aVARV-VK388: 100%, aVARV-VK470: 100%.

B) Genes inactivated in VARV, VARV-VD21, and certain or uncertain inactivations in some but not all aVARV sequences.

This includes five genes, *C9L* (29281), *B16R* (194427), *B12R* (191696), 193435*, and *A44L* (168244).

C9L: ANK/F-box protein related to host range [273], and an antagonist of the type I interferon response [372]. aVARV-VK388 and aVARV-VK470 share the same stop codons, some of which are shared with modern VARV. aVARV-VK382 has no coverage in the region that has a 2 nt insert that leads to the frameshift in aVARV-VK388 and aVARV-VK470, so the stop codons are different. Coverage of reference sequence (TATV): VARV-VD21: 58.5%, aVARV-VK382: 74.3%, aVARV-VK388: 91.7%, aVARV-VK470: 92.1%.

B16R: IL1β receptor [270]. Inactive in VACV-COP, but active in VACV-WR, where it is referred to as *B15R*. *'Deletion of B15R from vaccinia virus accelerated the appearance of symptoms of illness and mortality in intranasally infected mice, suggesting that the blockade of IL-1β by vaccinia virus can diminish the systemic acute phase response to infection and modulate the severity of the disease.' [296].* VARV and VARV-VD21 share the same gene-inactivating mutation. The aVARV sequences have different initial gene-inactivating mutations, but some of the downstream stop codons

agree with VARV. Coverage of reference sequence (TATV): VARV-VD21: 100%, aVARV-VK382: 78.9%, aVARV-VK388: 100%, aVARV-VK470: 100%.

B12R: Has 33% amino acid similarity to *B1*, a serine/threonine kinase. However, *B12* lacks enzymatic activity [373]. There is no coverage for aVARV-VK382. aVARV-VK388, aVARV-VK470, and VARV-VD21 are closest to modern VARV, based on sequence similarity, and have the same gene-inactivating mutation, caused by a 1 nt insertion at position 238 of modern VARV-BRZ, -SLN, -SAF, and -KUW. Coverage of reference sequence (VARV-SAF): VARV-VD21: 100%, aVARV-VK388: 100%, aVARV-VK470: 85.2%.

193435*: pCPXV0030. Surface glycoprotein [270]. VARV-VD21 and VARV have the same gene-inactivating mutation. The gene-inactivating mutations in aVARV-VK388 and TATV agree; some of the mutations are also present in VARV. aVARV-VK382 is most similar to CPXV in terms of insertions and deletions, but we have no coverage in the second half of the CPXV gene, so it is impossible to determine if there is a gene-inactivating mutation towards the C-terminus of the gene. There is no coverage for aVARV-VK470. Coverage of reference sequence (VARV-SLN): VARV-VD21: 100%, aVARV-VK382: 57.1%, aVARV-VK388: 92.9%.

A44L: 3-beta-hydroxysteroid dehydrogenase [275]. Deletion of *A44L* in VACV-WR does not affect viral replication in CV-1 cells in vitro, but the virus is attenuated in a mouse intranasal model [275]. *A44* synthesizes steroid hormones, induces immunosuppression, and contributes to the virulence of VACV-WR in mice [374, 375]. This gene is disrupted in all sequenced VARV hitherto. All aVARV sequences share the same stop codon. In aVARV-VK382, the stop codon is only covered by one read and is therefore listed as ‘uncertain’. Coverage of reference sequence (CMLV): VARV-VD21: 100%, aVARV-VK382: 94%, aVARV-VK388: 100%, aVARV-VK470: 100%.

C) Genes inactivated in VARV and VARV-VD21, but present in all aVARV sequences.

This includes four genes, *E7R* (68913), *B6R* (186591), *C5L* (33086), and *B7R* (187178).

E7R: EEV myristylated soluble protein [376]. Inactivation does not affect known biological properties of the virus [377]. Coverage of reference sequence (TATV): aVARV-VD21: 100%, aVARV-VK382: 88%, aVARV-VK388: 100%, aVARV-VK470: 100%.

B6R: Ankyrin [270]. Coverage of reference sequence (TATV): aVARV-VD21: 100%, aVARV-VK382: 90%, aVARV-VK388: 100%, aVARV-VK470: 100%.

C5L: Genome uncoating and DNA replication factor [372]. Coverage of reference sequence (CPXV-Gri): aVARV-VD21: 100%, aVARV-VK382: 88%, aVARV-VK388: 100%, aVARV-VK470: 100%.

B7R: is an endoplasmic reticulum-associated, putative chemokine binding protein. Virulence factor in VACV-WR [378]. Its deletion leads to the attenuation (smaller lesions) of VACV-WR in a murine intradermal model, but not in the murine intranasal model [378]. Coverage of reference sequence (TATV): aVARV-VD21: 100%, aVARV-VK382: 61%, aVARV-VK388: 100%, aVARV-VK470: 99% (the first two amino acids do not have coverage).

D) Genes inactivated in VARV and VARV-VD21, but present in some but not all aVARV sequences.

This includes nine genes, 22682, *A40R* (165942), 215231, *A57R* (180331), *C2L* (35634), *K1L* (41155), *F3L* (47695), 168145, and *A35R* (161464).

22682: CPXV020. Unknown [270]. VARV-VD21, aVARV-VK382, and aVARV-VK388 share the same stop codon. aVARV-VK470 is functional, and has reads covering 100% of the CMLV reference genome. Coverage of reference sequence (CMLV): VARV-VD21: 100%, aVARV-VK382: 65.5%, aVARV-VK388: 95.1%, aVARV-VK470: 100%.

A40R: Surface glycoprotein, immune modulator, related to C-type lectin-like proteins. Disruption of the *A40R* gene in VACV did not affect virus plaque size, in vitro growth rate and titre, EEV formation, or virus virulence in a murine intranasal model, but led to smaller lesions in the mouse intradermal model [379, 380]. Coverage of reference sequence (CPXV-Gri): VARV-VD21: 100%, aVARV-VK382: 66%, aVARV-VK388: 100%, aVARV-VK470: 39%. A novel stop codon is present in aVARV-VK388.

215231: pCPXV0002. N-methly D-aspartate receptor-like protein [270]. Gene-inactivating mutations agree in aVARV-VK388 and aVARV-VK470, present in aVARV-VK382. Coverage of reference sequence (CMLV): aVARV-VK382: 72.8%, aVARV-VK388: 100%, aVARV-VK470: 100%.

A57R: Guanylate kinase. An inactive gene in VACV-Cop and VACV-WR, with the same gene-inactivating mutation in mVARVs [274]. *A57* is predicted to only be func-

tional if it has an additional 5' region that contains an ATP binding domain [381]. In VARV, VARV-VD21, VACV-COP, VACV-WR, VACV-MVA, and RPXV, the gene is inactivated by an 11 nt deletion relative to CPXV. In HPXV, ECTV, TATV, CMLV, MPXV-ZAI, MPXV-WR, and aVARV-VK470, the gene is inactivated by nucleotide substitutions leading to a stop codon, or 1 nt insertions or deletions that cause a frameshift. The mutation that breaks MPXV-ZAI and MPXV-WR is identical, and the deletion that breaks aVARV-VK470 is also present in ECTV, although ECTV breaks earlier due to a substitution. CPXV-Ger, CPXV-Gri, CPXV-BR, aVARV-VK382, and aVARV-VK388 all contain the 5' region with the ATP binding domain. Coverage of reference sequence (CPXV-Gri): VARV-VD21: 100%, aVARV-VK382: 78%, aVARV-VK388: 100%, aVARV-VK470: 100%.

C2L: Kelch-like proteins with putative host-range function [382]. Deletion of *C2L* in VACV has no effect on growth in vitro, but has altered plaque morphology [297]. Deletion of *C2L* in VACV also does not affect virulence in a mouse intranasal model, but leads to larger lesions and more cell infiltration in a mouse intradermal model [297]. Thus, *C2L* contributes to cell cytopathic effect and reduces immunopathology in vitro [297]. Coverage of reference sequence (CPXV-Ger): VARV-VD21: 100%, aVARV-VK382: 85%, aVARV-VK388: 100%, aVARV-VK470: 100%. A novel stop codon is present in aVARV-VK470.

K1L: NFκB inhibitor, and human SAMD9 inhibitor [291, 383]. Plays a role in antiviral response and inflammation, and is a known host-range gene [273] since its deletion in VACV-COP results in a replication defect in rabbit RK13 cells [293]. The presence of a functional *K1L* gene in aVARV-VK382 and aVARV-VK388, as well as the presence of different gene-inactivating mutations in CMLV, TATV, and VARV [273] suggest that the gene was lost independently in CMLV, TATV, and VARV after the divergence of the aVARV clade and mVARV ~1.6 kya. Coverage of reference sequence (CPXV-Gri): VARV-VD21: 100%, aVARV-VK382: 86%, aVARV-VK388: 100%, aVARV-VK470: 100%. A novel stop codon is present in aVARV-VK470.

F3L: Kelch-like protein [270]. Deletion of *F3L* in VACV leads to smaller lesions in a mouse intradermal model, as indicated by smaller lesions, and alterations in the innate immune response [299]. Coverage of reference sequence (TATV): VARV-VD21: 100%, aVARV-VK382: 93%, aVARV-VK388: 100%, aVARV-VK470: 100%. A novel stop codon is present in aVARV-VK470.

168145: CPXV181. Unknown [270]. Coverage of reference sequence (CMLV): VARV-VD21: 100%, aVARV-VK382: 100%, aVARV-VK388: 100%, aVARV-VK470: 100%. A novel stop codon is present in aVARV-VK470.

A35R: Virulence factor [7, 270]. Inhibits MHC class II-restricted antigen presentation, immune priming of T lymphocytes, and subsequent chemokine and cytokine synthesis [384]. Coverage of reference sequence (RPXV): VARV-VD21: 100%, aVARV-VK382: 85.8%, aVARV-VK388: 100%, aVARV-VK470: 100%. A novel stop codon is present in aVARV-VK382.

E) Genes inactivated in VARV and VARV-VD21, and uncertain inactivations in aVARV sequences.

This includes three genes, *A52R* (175891), *A55R* (178726), and 214439.

A52R: Intracellular Bcl2-like protein, blocks NFκB activation, and contributes to virulence in VACV [385–387]. Deletion of *A52R* from VACV leads to attenuation in a murine intranasal model [385]. The uncertain stop codon is identical between all aVARV sequences. The stop codon in VARV-VD21 is novel. Coverage of reference sequence (CPXV-Ger): VARV-VD21: 100%, aVARV-VK382: 89.8%, aVARV-VK388: 100%, aVARV-VK470: 100%.

A55R: Kelch-like protein, immunomodulator [7, 270]. Deletion of *A55R* in VACV-WR does not affect growth rate in vitro but leads to a different plaque morphology and cytopathic effect. It also leads to larger lesions in a murine intradermal model, hence affecting the host response to infection in vivo [298]. Coverage of reference sequence (CPXV-Gri): VARV-VD21: 53.2%, aVARV-VK382: 85%, aVARV-VK388: 100%, aVARV-VK470: 99.7%. An uncertain novel stop codon in aVARV-VK470, covered by two reads.

214439: pCPXV0002. *CrmE*. Tumor necrosis factor (TNF) receptor homolog. *CrmE* inhibits the cytotoxic and apoptotic activities of human, but not mouse or rat TNF in vitro. In a murine intranasal model, VACV-USSR recombinants lacking *CrmE* are attenuated. Expression of *CrmE* in VACV-WR enhances virulence in the murine model [370]. Absent in both VARV-VD21 and modern VARV. A novel stop codon is present in aVARV-VK382, at nucleotide position 70 relative to CPXV-Gri, but only has coverage of one read. The gene appears to be functional in aVARV-VK388 and aVARV-VK470. *CrmE* sequences present in aVARV-VK388 and aVARV-VK470 are most closely related to CPXV-Gri and CPXV-Ger. Coverage of reference sequence (CPXV-Ger): aVARV-VK382: 97.6%, aVARV-VK388: 100%, aVARV-VK470: 100%.

Active *CrmE* orthologs are only found in CPXV-Gri, CPXV-Ger, VACV-LIS, VACV-Lc16m8, VACV-Lc16mO, and VACV-USSR strains, but are absent in all other VACV strains as well as in HSPV, RPXV, VARV, and TATV, and are truncated in CMLV and ECTV [273]. Gene-inactivating mutations are present in identical positions in MPXV, ECTV, and CPXV-BR, suggesting that the inactivation arose either in an ancestral virus, or as a result of horizontal gene transfer or recombination events [273], in which case aVARV-VK388 and aVARV-VK470 may have also received their version of *CrmE* in a similar fashion.

F) Genes present in VARV and VARV-VD21, but absent in some aVARV sequences.

This includes two genes, *E5R* (66488) and *C1L* (37242).

E5R: Virosome component protein [270, 388]. Functional in aVARV-VK388. Novel identical stop codons in aVARV-VK382 and aVARV-VK470. The stop codon in aVARV-VK382 is listed as uncertain, since read coverage is below four. Coverage of reference sequence (CMLV): VARV-VD21: 100%, aVARV-VK382: 94.8%, aVARV-VK388, and aVARV-VK470: 100%.

C1L: Unknown. Bcl-2-like [270]. Coverage of reference sequence (VARV-KUW): VARV-VD21: 100%, aVARV-VK382: 85%, aVARV-VK388: 99.1%, aVARV-VK470: 100%. A novel stop codon is present in aVARV-VK388.

G) Genes present in VARV and VARV-VD21, but with uncertain inactivations in some aVARV sequences.

This includes 17 genes, *C10L* (23924), *A49R* (172501), *B18R* (197359), *A47L* (170515), *B17L* (194473), *B20R* (200936), *A10L* (135433), *E9L* (69869), *J6R* (103459), *D1R* (112743), *A3L* (127555), *A12L* (139081), *A18R* (143956), *A22R* (146293), *A24R* (150952), *A33R* (160360), and *A36R* (162202).

A number of genes in this set are related to virus replication, so their putative absence should be treated with caution. The *D1R* gene is included in this category despite having five reads that indicate a gene-inactivating mutation. Because the mutation is due to a G $\rightarrow$ A change and the reads have identical start and end positions, it is assumed that the reads are PCR duplicates, and that the apparent mutation is actually DNA damage.

H) Genes absent or no coverage in VARV, VARV-VD21, and aVARV sequences, and with gene-inactivating mutations in CMLV or TATV.

This includes three genes, *K6L;K5L* (45027), 202819, and *B10R* (190433).

K6L;K5L: Monoglyceride lipase (putative); lysophospholipase-like protein [270].

202819: CPXV216, VACV-WR-B13R. *CrmA*, serine protease inhibitor. Inhibits both extrinsic apoptosis and the host inflammatory response [7, 389, 390]. Deletion of *CrmA* has no effect [391] or leads to attenuation of VACV-WR in the intranasal murine model, and leads to larger skin lesions in the intradermal murine model [380].

B10R: Unknown [270].

APPENDIX B: DEVELOPING PARAMETERS AND METHODS FOR THE LIGHT MATTER ALGORITHM

In order to create the light matter algorithm described in the second part of this thesis, the following components had to be implemented: the features that are identified by the algorithm, the significance methods used to assess if the histogram peaks constitute significant matches, and methods for scoring the matches that were identified as significant. Also, tools were needed to evaluate and visualise the different components of the algorithm. In the process of implementing these components, a high number decisions had to be made. In some cases, these were made based on outside evidence, sometimes different alternatives were tested, and in other cases the seemingly most straight-forward option was chosen. In this chapter, I describe these components, starting with the methods for evaluation and visualisation, since those will be used in later parts of the chapter, and then move on to describe the development of features, significance methods, and scoring methods.

B.1 TOOLS TO EVALUATE THE LIGHT MATTER ALGORITHM

The development and optimisation of the light matter algorithm requires tools to evaluate its results. Given that the algorithm is built to compare sequences based on predicted structural features, it makes sense to evaluate its performance using pairs of sequences for which the structural and sequence similarity is known. The protein data bank (PDB) provides such information. Founded in 1971, the PDB is the primary repository for three-dimensional structural data on proteins and other biological macromolecules [362]. Since 2003, the repository has been managed by an international consortium, the world-wide Protein Data Bank, whose partners include the Research Collaboratory for Structural Bioinformatics, the Macromolecular Structure Database at the European Bioinformatics Institute, the Protein Data Bank Japan, and others [392]. The information stored in PDB includes the protein name, atomic coordinates of the protein structure, the sequence, authors, key references, and also some derived data, such as quality assessment and fold classification [347].
Using the structure and sequence information from PDB, two areas of the light matter algorithm can be evaluated: firstly, the identification of features based on secondary structures, and secondly, the scores that are assigned to a match and how they corre-

spond to the structural reality. Furthermore, visualisations of the location and pairing of features on a sequence and on a three-dimensional structure help to improve our understanding of the behaviour of the algorithm.

B.1.1 Evaluation using perfect finders

Identification of features that can meaningfully detect similarity between sequences is a crucial step in the light matter algorithm. While some features can be identified unambiguously (types of amino acids that are highly conserved, features based on amino acid properties, and sequence motifs), the identification of features based on secondary structures is more error-prone. In order to evaluate feature identification and performance of the light matter algorithm, knowing the actual locations of secondary structures within a sequence is necessary. We can use that information to evaluate two things: first, whether the feature finders designed to identify secondary structures are functioning correctly, and second, we can compare the behaviour of the light matter algorithm using the correct secondary structure information to its behaviour using the imperfect features based on secondary structures that are identified by the experimental landmark feature finders.

Apart from the coordinate files of the structure and information such as the sequence or the authors, each PDB entry also contains the type of secondary structure each amino acid in the amino acid sequence occurs in. The secondary structures are assigned using the DSSP algorithm, based on the positions of the atoms in the protein structure and by calculating hydrogen bond energy between all atoms [393]. Eight types of secondary structure elements are assigned letter codes as follows [393]: Alpha helix (H), Alpha helix 3-10 (G), Alpha helix pi (I), Residue in isolated beta bridge (part of a beta strand, B), Extended strand (participates in beta ladder, E), Hydrogen bonded turn (pieces of helix too short to be a helix, T), Bend (region of high curvature, S), and Coil (residues that are not in any of the above conformations, C). Thus, if a sequence is present in PDB, the secondary structure information assigned to it can be used to extract features which use the correctly-annotated secondary structure categories. For the purpose of evaluating the feature finders based on predicted secondary structures, and the performance of the light matter algorithm under correctly identified secondary structure features, we implemented five idealised (i.e., 'perfect') landmark finders based on the PDB entries. The landmark finders identify the alpha helix, alpha helix 3-10, alpha helix pi, and extended strand secondary structures, as well as a finder for all three alpha helix types combined. The finders were named

'PDB AlphaHelix', 'PDB AlphaHelix_3_10', 'PDB AlphaHelix_pi', 'PDB ExtendedStrand' and 'PDB AlphaHelix_combined'.

B.1.2 Evaluation using test datasets of sequences with known structural similarity

The scores that are generated by the light matter algorithm need to be biologically relevant. Given that the algorithm compares sequences based on predicted structural features, scores of sequence similarity (such as the BLAST bit score) do not provide enough resolution over large evolutionary distances to evaluate the light matter algorithm. Therefore, we use the correlation between measures of structural similarity between proteins and the scores obtained from the light matter algorithm for evaluation. The sequence and structural information from PDB allowed me to construct five test datasets of pairs of sequences where the structural similarity and sequence similarity are known.

I constructed five test datasets, referred to as '2HLA', '4MTP', 'Polymerase', 'HA', and '4PH0'. Two datasets consist of sequences of viral RNA dependent RNA polymerases (the 4MTP and Polymerase datasets), and one dataset each of sequences of the major histocompatibility complex (2HLA), influenza haemagglutinin (HA), and viral capsid proteins (4PH0). Structural similarity was measured using the Dali Z-score. The Dali algorithm computes the Z-score by comparing two distance matrices of distances between C-alpha atoms within a protein [360]. Sequence similarity was measured using the bit score reported by BLASTp. The bit score measures sequence similarity by scoring matches between sequences using amino acid replacement frequencies [394].

Dataset	Number of sequence pairs	Median sequence length
2HLA	141	275
4MTP	215	563
Polymerase	484	563
HA	169	328
4PH0	113	215

Table B.1: Number of sequence pairs and sequence lengths in each test dataset. Sequence lengths are given in amino acids.

The sequences in the Polymerase dataset are from Černỳ *et al.*, (2014) [343] and the sequences in the HA dataset were provided by David Burke. The 2HLA, 4MTP, and 4PH0 datasets were constructed by comparing the 2HLA:A, 4MTP:A, and 4PH0:A

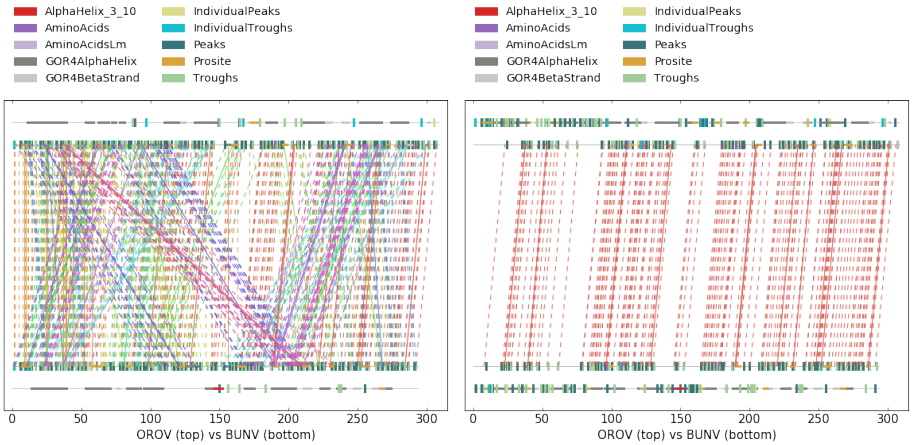
structures against all known structures using the online Dali server [395]. The resulting list of similar structures was subsampled, to achieve an even spread across the whole range of Dali Z-scores. In the 4PH0 dataset, I also removed all sequences that contained the words 'complex' or 'muta' in the title, in an attempt to only include structures that were not a in complex with a ligand or that had known mutations. The reduced lists of sequences from each dataset were compared against the 2HLA:A, 4MTP:A, and 4PH0:A sequence respectively, using the BLASTp algorithm [25] to acquire the BLAST bit scores as a measure of sequence similarity. For the Polymerase and HA datasets, pairwise comparisons using BLASTp [25] and the Dali [360] algorithm were performed to acquire the pairwise sequence and structural similarity scores. The Polymerase dataset is the largest test dataset, containing 484 distances, whereas the 4PH0 dataset is the smallest, with 113 distances (Table B.1).

B.1.3 Visualisations

Two types of visualisations were implemented and used in the process of developing the light matter algorithm, to improve my understand its behaviour:

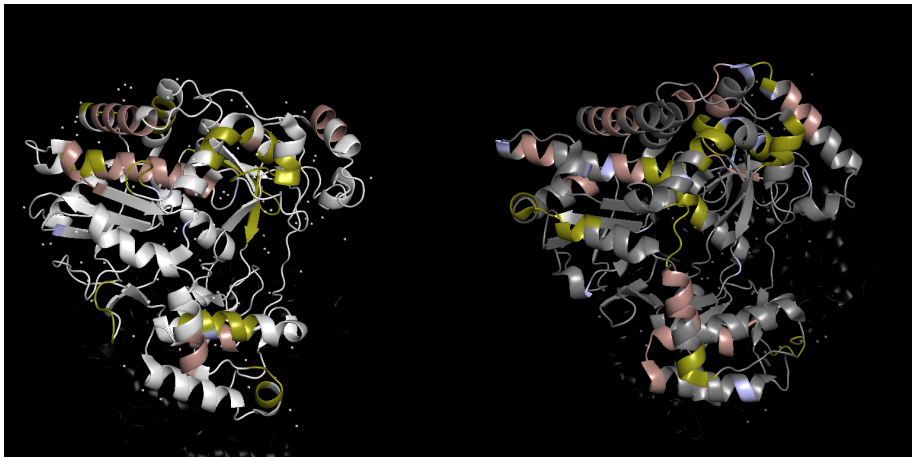
Terry Jones developed the Horizontal line plot, shown in Fig. B.1. The horizontal line plot shows how the features are positioned and the way they are paired and then binned. The two horizontal lines at the bottom and top of the figure indicate the subject and the query sequences. The small, bold, coloured horizontal lines plotted over the horizontal lines indicate the landmarks, while the vertical lines indicate the trig points, where each colour corresponds to a feature type, as per the figure legend. Features that are involved in a match are shown in the two innermost lines, whereas features not involved in a match are shown on the two outermost lines. The sloping lines between the query and the subject connect the matching features. A solid line denotes matching landmarks, and a dotted line indicates matching trig points. Matching features that are in the same bin are connected with lines with the same colour. Variations of the plot exist, where the significant bins only (Fig. B.1a), the best bin only (Fig. B.1b), or all bins are shown.

I developed visualisations to get a visual impression of the three-dimensional structure and of the location of features on the structure. An example is shown in Fig. B.2. The structures are displayed in PyMOL [396] using the coordinate files stored in PDB. It is possible to display multiple structures next to each other, to display different features on each of them, to annotate the features that are involved in a match, and to display lines between the features that match between two structures.

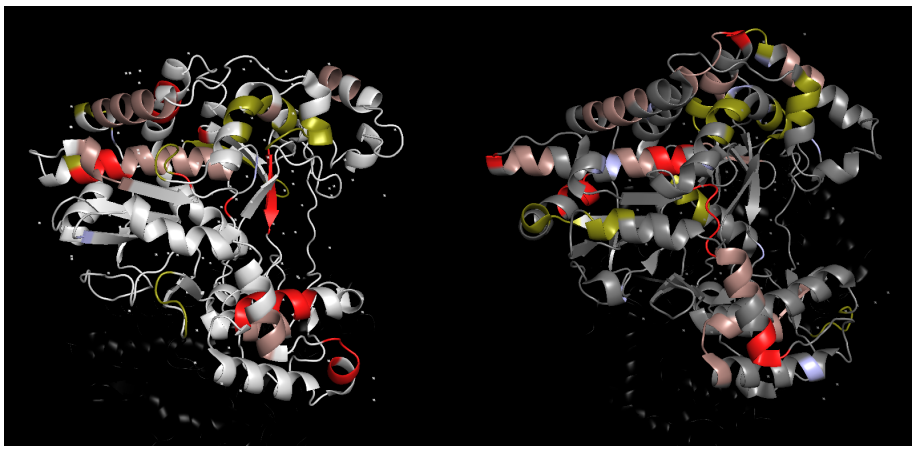


(a) **(b)**

Figure B.1: Horizontal line plot. Subject and query sequences are shown as horizontal lines at the bottom and top of the figure. Landmarks are shown as coloured lines and trig points as coloured bars. Matching features are indicated on the inner line, and non-matching features on the outer line. Matching features in the subject and the query are connected by lines. If the matching features are in the same histogram bin, the connecting line is in the same colour. textbfa) only the significant bins are shown, **b)** only the best bin is shown.



(a)



(b)

Figure B.2: Viral RNA dependent RNA polymerase structures displayed in PyMOL. Left: 1XR7:A, right: 1KHV:B. **a)** Features found by the 'AC_AlphaHelix' feature finder are shown in brown, features found by the 'AC_ExtendedStrand' feature finder in green, and features found by the 'AminoAcidsLm' feature finder in mauve. **b)** Same structures and colouring as in Fig. a), but features in pairs that match are shown in red.

B.2 FEATURES

Sequence comparisons can be performed using different alphabets. Current sequence matching algorithms typically use alphabets based on nucleotides or amino acids [26, 331]. The light matter algorithm compares sequences based on an alphabet of features that are expected to be structurally conserved. In order to be effective, the features used in the light matter algorithm must fulfil the following criteria:

1. Feature detection must be independent of position in the sequence, to allow the matching of features in sequence fragments to those in complete sequences.
2. Features need to be robust to distortion from mutations or sequencing errors.
3. Features must be small enough to identify in a sequence of around 65 amino acids, due to the short length of reads generated by most Next Generation Sequencing (NGS) technologies.
4. Features need to be structurally relevant. This does not mean that they need to correspond to an exact part of the structure, but they need to carry some information relevant for the formation of the structure.

Features are divided into two groups, 'landmarks' and 'trigonometry (trig) points', and are identified by so called 'finders'. Landmarks can form pairs with other landmarks or trig points, but trig points cannot form pairs between each other. The division of features into landmarks and trig points facilitates the pairing of features. The description of the Shazam algorithm only provides a rudimentary description of which features can form pairs with each other [355]. Splitting the features into two groups, only one of which can initiate the pairing (the landmarks), is a convenient way for us to choose those features. The light matter algorithm implements landmark finders based on seven types of structural features that are encoded in the amino acid sequence. All are expected to be relevant for the shape of the protein structure: alpha helices, including alpha helix 3-10 and alpha helix pi, beta strands, coils, conserved amino acids such as cysteine and tryptophan, PROSITE motifs and short linear motifs. All landmark finders available to the algorithm are shown in table B.2. Trig points are based on less obvious patterns in amino acid properties and cannot readily be associated with a particular part of the structure, but may still be important for the formation and stability of a protein structure. These may include features such as hydrophobicity patterns or other patterns based on amino acid properties. All trig points available to the algorithm are shown in table B.3. The following sections describe how the different feature finders were implemented.

Landmark name	Description
AC AlphaHelix	Identifies alpha helices based on substrings marked as alpha helix in PDB (chapter B.2.1.1).
AC AlphaHelix_3_10	Identifies alpha helix 3-10 based on substrings marked as alpha helix 3-10 in PDB (chapter B.2.1.1).
AC AlphaHelix_pi	Identifies alpha helix pi based on substrings marked as alpha helix pi in PDB (chapter B.2.1.1).
AC AlphaHelix_combined	Identifies alpha helices based on substrings marked as alpha helix, alpha helix 3-10 or alpha helix pi in PDB (chapter B.2.1.1).
AC ExtendedStrand	Identifies extended strands based on substrings marked as extended strand in PDB (chapter B.2.1.1).
AlphaHelix	Identifies alpha helices based on the assumption that an alpha helix is composed of at least one repeat of one hydrophobic followed by three hydrophilic amino acids (chapter B.2.1.1).
AlphaHelix_3_10	Identifies alpha helix 3-10 based on the assumption that an alpha helix 3-10 is composed of at least one repeat of one hydrophobic followed by two hydrophilic amino acids (chapter B.2.1.1).
AlphaHelix_pi	Identifies alpha helix pi based on the assumption that an alpha helix pi is composed of at least one repeat of one hydrophobic followed by four hydrophilic amino acids (chapter B.2.1.1).
BetaStrand	Identifies beta strands by screening for a sequence of at least six consecutive amino acids from the set [V, I, C, F, Y, T] [348] (chapter B.2.1.1).
BetaTurn	Consists of four amino acids. At each position the amino acid must come from the sequence of sets [N, C, D], [P, S, K], [N, D, G], [W, G, Y] [397] (chapter B.2.1.1).
AminoAcidsLm	Identifies cysteines (chapter B.2.1.3)
GOR4AlphaHelix	Identifies alpha helices using the GOR4 algorithm [398] (chapter B.2.1.1).
GOR4BetaStrand	Identifies beta strands using the GOR4 algorithm [398] (chapter B.2.1.1).
GOR4Coil	Identifies coils using the GOR4 algorithm [398] (chapter B.2.1.1).
Prosite	Identifies patterns from the prosite database [349] (chapter B.2.1.2).
EukaryoticLinearMotif	Identifies eukaryotic linear motifs from the ELM database [352] (chapter B.2.1.2).
PDB AlphaHelix	Identifies alpha helices based on the assignment of alpha helices in PDB (chapter B.1.1). For evaluation only.
PDB AlphaHelix_3_10	Identifies alpha helix 3-10 based on the assignment of alpha helix 3-10 in PDB (chapter B.1.1). For evaluation only.
PDB AlphaHelix_pi	Identifies alpha helix pi based on the assignment of alpha helix pi in PDB (chapter B.1.1). For evaluation only.
PDB ExtendedStrand	Identifies beta strands based on the assignment of extended strands in PDB (chapter B.1.1). For evaluation only.
PDB AlphaHelix_combined	Identifies alpha helices based on the assignment of alpha helices, alpha helix 3-10 and alpha helix pi in PDB (chapter B.1.1). For evaluation only.

Table B.2: Description of the different landmark feature finders that are used by the light matter algorithm. Where relevant, chapter numbers are given.

Trig point name	Description
IndividualPeak	Based on a window of length three moved along the amino acid sequence. If the property values for composition, isoelectric point, and polarity all are higher for the current amino acid than for the two amino acids on either side, the current amino acid is an IndividualPeak (chapter B.2.2.2).
IndividualTrough	Based on a window of length three moved along the amino acid sequence. If the property values for composition, isoelectric point, and polarity are all lower for the current amino acid than for the two amino acids on either side, the current amino acid is an IndividualTrough (chapter B.2.2.2).
Peak	Based on a window of length three moved along the amino acid sequence. If the sum of the property values for composition, isoelectric point, and polarity is higher for the current amino acid than for the two amino acids on either side, the current amino acid is a Peak (chapter B.2.2.2).
Trough	Based on a window of length three moved along the amino acid sequence. If the sum of the property values for composition, isoelectric point, and polarity is lower for the current amino acid than for the two amino acids on either side, the current amino acid is a Trough (chapter B.2.2.1).
AminoAcids	Tryptophans (chapter B.2.2)

Table B.3: Description of the different trig point feature finders that are used by the light matter algorithm. Where relevant, chapter numbers are given.

B.2.1 Landmarks

This sub-chapter describes the development of the landmark finders. First, I will introduce the landmark finders based on secondary structures, then the landmark finders based on sequence motifs, and finally the one based on amino acids.

B.2.1.1 Landmarks based on secondary structures

Alpha helices and beta strands are integral parts of the structure of a protein [399] and are often highly conserved [400]. This makes them an obvious choice for landmarks in the light matter algorithm. Secondary structures can be identified using secondary structure prediction tools, with the best and most modern tools reaching an accuracy of about 80% [347]. The first secondary structure prediction algorithms relied on statistical analysis of sequence composition within the different types of secondary structures. They calculated the probability of each amino acid to be in a particular secondary structure, also called amino acid propensities [401]. Examples include the Chou-Fasman method [402] and the GOR algorithm [403]. The accuracy of these early prediction methods was between 50% and 60% [404]. Those methods were improved by calculating amino acid propensities for a window around the residue of which the secondary structure should be predicted, rather than considering each residue on its own [401]. Further improvements were made by incorporating multiple sequence alignments and machine learning approaches, as well as by the availability of a higher number of three-dimensional structures due to the growth of public structure databases [347]. Sequences related to the query sequence are identified using BLAST, PSI-BLAST, or hidden Markov models and converted to a multiple sequence alignment or a position specific scoring matrix. This information is used as the input to a machine learning method, such as a neural network, k-nearest-neighbour approach, another hidden Markov model, or a combination thereof [347]. Neural networks generally have an input layer, a hidden layer, and an output layer. The network takes an input, in our case a sequence or a sequence profile, and calculates the output value, which in this case is the secondary structure element of a residue, based on a set of functions with associated weights in the hidden layer. Tools using this approach include PHD [405], PSIPRED [406], and SSPro [407], among many others [347].

For the purpose of the light matter algorithm, secondary structure prediction needs to be fast and cannot rely on sequence similarity to known sequences, thus ruling out methods incorporating multiple sequence alignments. This limits us to the earlier,

less-accurate secondary structure prediction methods, such as the first four versions of the GOR algorithm [403].

Basic secondary structure finders

Initially, we implemented finders that identify alpha helices based on patterns of hydrophobic and hydrophilic amino acids. Beta strands are identified by screening for a sequence of at least six specific amino acids, while the beta turns have to consist of a sequence of four amino acids, where at each of the four positions, an amino acid from a certain set has to occur (see table B.2). The finders are called 'AlphaHelix', 'AlphaHelix_3_10', 'AlphaHelix_pi', 'BetaStrand' and 'BetaTurn'. *The alpha helix and beta strand finders were written by Terry Jones.* Table B.4 shows performance statistics of the five basic secondary structure finders when evaluating them against the secondary structure annotations for all sequences in PDB, downloaded on 11 July 2016. The low sensitivity for all five finders suggests that they miss a high proportion of the respective secondary structures present in PDB. The low precision for the three alpha helix finders also suggests that they identify a high number of false positives. These results suggest that the features identified by the five basic secondary structure finders do not contribute to the algorithm's performance and should be replaced by more sophisticated finders, in the interest of computation speed of the algorithm.

	AlphaHelix	AlphaHelix_3_10	AlphaHelix_pi	BetaStrand	BetaTurn
True positive	138,053	139,168	0	44,903	63,583
True negative	45,286,095	61,165,671	66,180,920	51,762,384	53,567,407
False positive	474,687	2,726,606	114,089	26,009	75,593
False negative	20,409,621	2,277,011	13,447	14,475,160	12,601,873
Sensitivity	0.01	0.06	0.00	0.00	0.01
Precision	0.23	0.05	0	0.63	0.46

Table B.4: Performance statistics of the basic secondary structure finders when evaluated using the secondary structure annotations in PDB. The sensitivity is calculated as (true positives / (true positives + false negatives)) and the precision as (true positives / (true positives + false positives)). The sequences from PDB were downloaded on 11 July 2016.

GOR4 secondary structure finders

In order to improve the identification of secondary structures, we next implemented finders that rely on a dedicated secondary structure prediction algorithm. As mentioned earlier, for reasons of speed, and since the overall aim is to identify query sequences without sequence similarity to known sequences, we had to rely on a secondary structure prediction algorithm that does not require a multiple sequence alignment. The GOR method for secondary structure prediction was originally developed in 1978 by Garnier, Osguthorpe, and Robson [403]. It has since gone through various refinements, ultimately reaching an accuracy of around 73% in its 5th version [408]. We used the 4th version of the GOR algorithm [398] (GOR4), as it is the most advanced version of the algorithm that does not require a multiple sequence alignment. The algorithm has a published accuracy of 64.4% [398]. *The Python wrapper of the GOR4 algorithm was written by Terry Jones.*

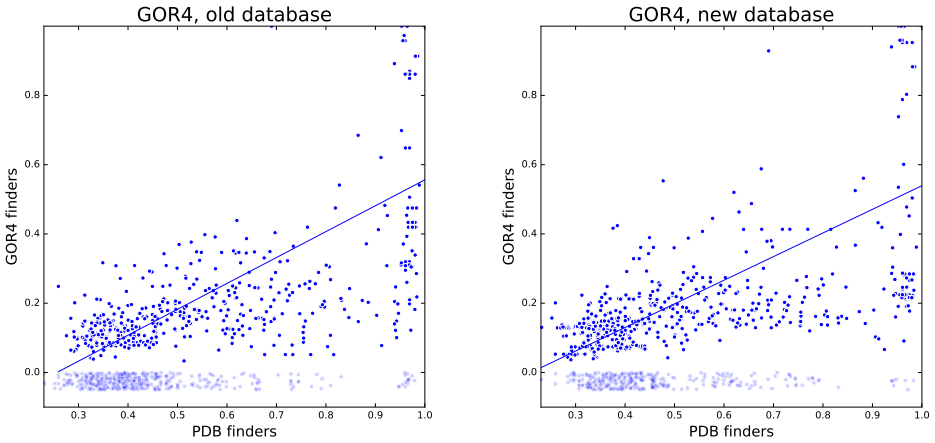
The database which the GOR4 algorithm uses by default contains 267 sequences, reflecting the number of structures in PDB around 1996, when the algorithm was developed. However, given that PDB has expanded since then, I updated the GOR4 database with all sequences in PDB up to 2016, to increase prediction accuracy. To construct the new database, I followed the steps that were taken to build the original database, outlined in Garnier *et al.,* (1996) as closely as possible [398][1]. All sequences and associated structures present in PDB on 3 March 2016 were filtered for length >50 amino acids and for a resolution <2.5Å. The remaining sequences were clustered using CD-HIT [409] with a sequence identity cut-off of 30%. For the

[1] Apart from the steps explained in the following text, Garnier *et al.,* (2016) also excluded sequences with an R-factor bigger than 25%. Since this information was not easily accessible, I omitted this step.

resulting 13,329 sequences, the secondary structures assigned to each amino acid in PDB using DSSP had to be translated from the eight-letter assignment to the three-letter assignment that is used by GOR4. Finally, the sequences were formatted to be useable by the GOR4 algorithm. Garnier *et al.,* (1996) [398] test the accuracy of the GOR4 algorithm using a jackknife test, leaving out one sequence from the database and predicting its secondary structure from the remaining sequences in the database. I did the same, once for the 267 sequences from the original GOR4 database and once for the new database consisting of 13,329 sequences. I could not reproduce the overall accuracy of 64.4% that Garnier *et al.,* (2016) [398] report using the old database of 267 sequences, but instead calculated an accuracy of 62.7%. The cause of this discrepancy may stem from using a slightly different version of the GOR4 algorithm than the one used in the original publication, since the code that is available online was last modified in 1997. The accuracy of the GOR4 algorithm when using the new database with 13,329 sequences is 66.4%. When evaluating the old and new GOR4 databases against the whole PDB database downloaded on 11 July 2016, I found a sensitivity and precision of 0.598 and 0.453 for the old database and 0.606 and 0.459 for the new database, respectively. I used the five test datasets described in chapter B.1.2 to investigate the correspondence between scores obtained using the perfect finders (described in chapter B.1.1) and scores obtained using the old and new GOR4 databases, respectively (Fig. B.3). Using the old GOR4 database, the scores from using the GOR4 finders correspond slightly better to the scores obtained from the perfect PDB finders, but the difference between the two regression lines is not significant (P value: 0.067). There was no negative effect on runtime when using the larger database. Given the higher accuracy and slightly higher sensitivity and precision of the GOR4 algorithm when using the new database, and since there is no negative effect on the speed of the algorithm, I decided to use the updated version of the GOR4 database in the light matter algorithm.

Aho-Corasick finders

Secondary structure prediction programs that are able to execute in a useful time-frame for our purposes, such as the GOR4 algorithm, have an accuracy of around 65% [347, 398]. Thus, their suitability for the light matter algorithm may be limited. A feature finder that is able to identify secondary structures with a high precision may be more desirable, even if this also leads to a high number of false positives. Searching for substrings of amino acid sequences that are known to occur frequently in certain secondary structures would always identify their corresponding secondary structures correctly, as well as making some false positive identifications.



(a) **(b)**

Figure B.3: Correspondence between scores obtained using the perfect PDB finders and scores obtained using the GOR4 finders using the a) old and b) new GOR4 databases. The matches come from the five test datasets described in chapter B.1.2. Scores using the perfect PDB finders and GOR4 finders are shown on the x-axis and y-axis, respectively. The blue line shows the linear regression and 95% confidence interval. Matches that have a score of 0.0 using the GOR4 finders (shown as pale dots), were jittered in the y dimension to improve readability, and were not included in the linear regression. The slope and correlation coefficient for the old and new databases are 0.75, 0.7 (R^2=0.49) and 0.68, 0.66 (R^2=0.44), respectively. There is no significant difference between the regression lines in a) and b) (P value: 0.067). The 'GOR4AlphaHelix' and 'GOR4BetaStrand' finders and the 'PDB AlphaHelix', 'PDB AlphaHelix_3_10', 'PDB AlphaHelix_pi', and 'PDB ExtendedStrand' finders were used.

The following sections describe the implementation of landmark finders that are based on the secondary structure assignments in PDB. The resulting landmark finder scans an amino acid sequence for the presence of substrings that are known to occur frequently in a particular secondary structure.

Using the secondary structure assignments in PDB, downloaded on 11 July 2016, I extracted all strings of amino acid sequences that occur in a certain secondary structure element longer than four amino acids, as well as all possible substrings thereof, also longer than four amino acids. Five secondary structure elements were considered: alpha helix, alpha helix 3-10, alpha helix pi, extended strand, as well as the three types of alpha helix combined (referred to as 'alpha helix combined'). Only PDB structures with a resolution <3.0Å and non-obsolete entries were used. As it is impractical to use all resulting substrings due to computational limitations, I needed to select a subset of all substrings that were extracted. Ideally, the substrings in the

subset should occur often and have a high precision. In order to select substrings, each substring was matched to all sequences in PDB. If there was a match, and it occurred in the correct secondary structure element, the match was counted as a true positive (TP), if the match was in the wrong secondary structure element, it was counted as a false positive (FP). Multiple subsets of substrings were generated using different cut-offs for the absolute number of occurrences as well as the percentage of true positives. If multiple substrings were nested (such as 'AHGC' and 'AHGCA') and have the same number of true and false positives, only the shortest substring was included in the subset. The subsets of substrings were evaluated using the following criteria: first, the subset of substrings used in the finder should have a high precision averaged over all substrings in the subset. The precision was calculated using the following formula:

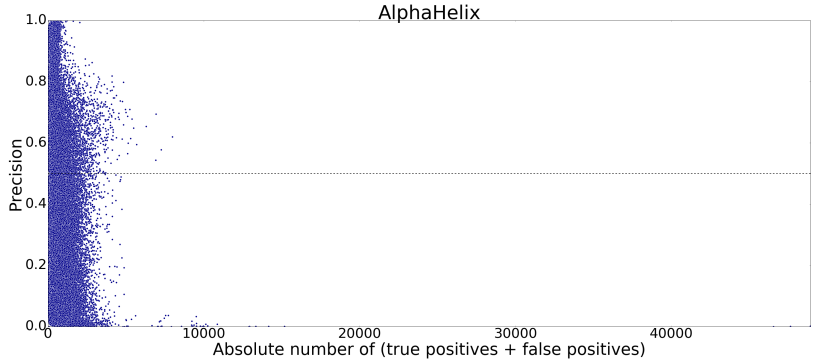
$$Precision = \frac{TP}{TP + FP} \tag{B.1}$$

Second, the substrings in the subset should match a high proportion of known secondary structures in PDB. Third, there should be a good correlation between the light matter scores calculated using the perfect finders described in chapter B.1.1 and the finders using the subsets of substrings. And finally, smaller subsets were preferred, so as to not negatively affect the speed of the algorithm. According to these criteria, a subset of substrings was selected for each of the five secondary structure elements.

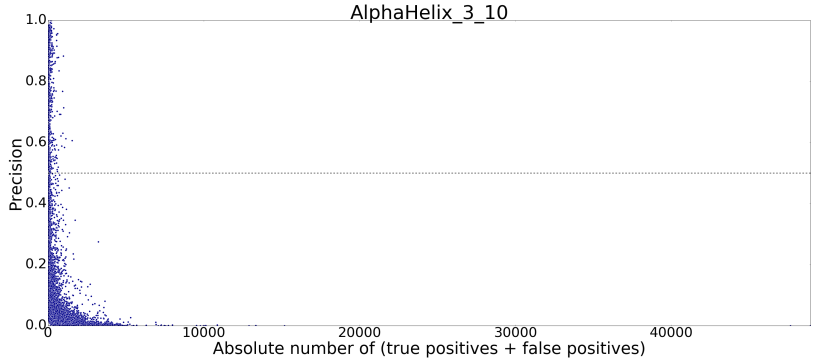
Table B.5 shows for each of the five secondary structure elements, the total number of structure strings, the number of unique structure strings, and the number of unique substrings. Substantially more unique substrings are available for AlphaHelix, AlphaHelix_combined, and ExtendedStrand.

Secondary structure element	Number of structure strings in PDB	Number of unique structure strings in PDB	Number of unique substrings
AlphaHelix	1,815,854	381,066	12,169,718
AlphaHelix_3_10	717,799	32,114	61,487
AlphaHelix_pi	2,518	552	2,030
AlphaHelix_combined	2,393,589	410,041	12,895,213
ExtendedStrand	2,548,920	276,463	1,772,181

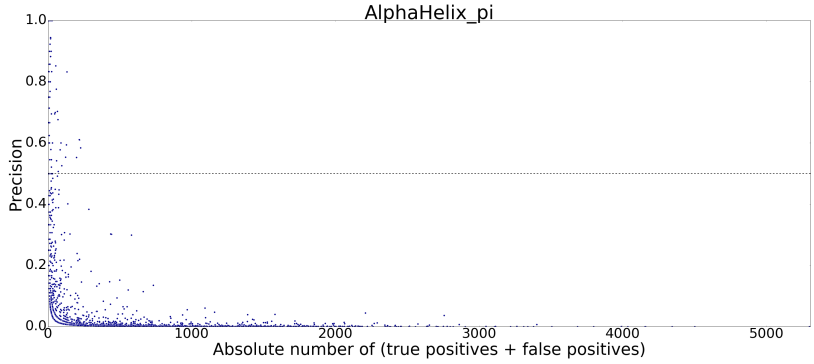
Table B.5: Number of structure strings and substrings in PDB.


AlphaHelix
Precision
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
30000
40000
Absolute number of (true positives + false positives)

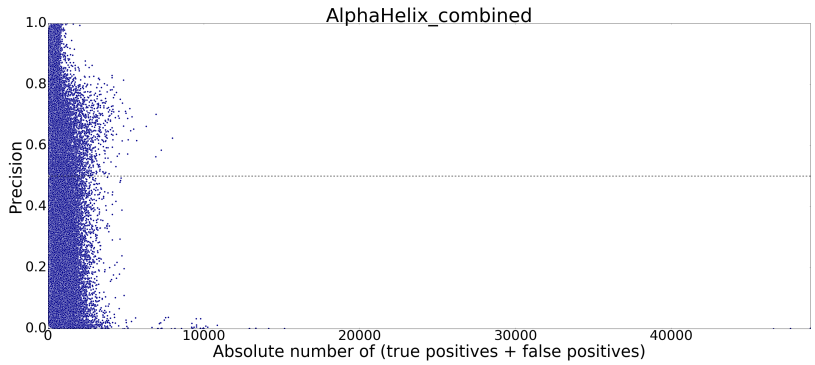
(a)


AlphaHelix_3_10
Precision
1.0
0.8
0.6
0.4
0.2
0.0
0
10000
20000
30000
40000
Absolute number of (true positives + false positives)

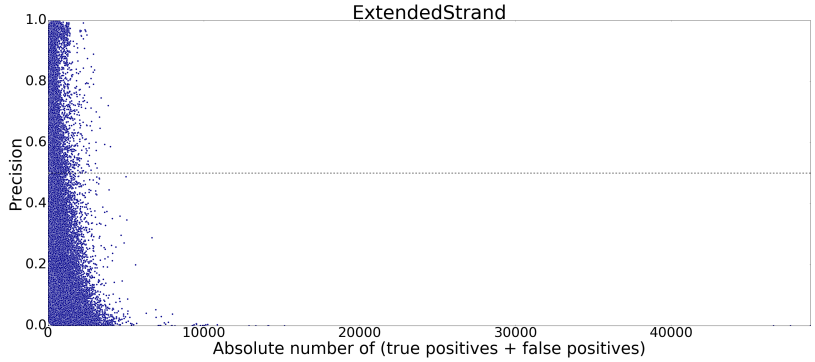
(b)


AlphaHelix_pi
Precision
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
5000
Absolute number of (true positives + false positives)

(c)



(d)



(e)

Figure B.4: Scatter plots of the precision (Percentage of true positives) against the total number of true and false positives for each substring. a) AlphaHelix, all lengths, b) AlphaHelix_3_10, all lengths, c) AlphaHelix_pi, all lengths, d) AlphaHelix_combined, all lengths, e) Extended-Strand, all lengths. Each plot corresponds to a different secondary structure element.

Figures B.4a–e) show for each secondary structure element, each substring plotted by its precision on the y-axis and the sum of its true and false positives evaluated using PDB (which is equal to the number of times the substring occurs in PDB) on the x-axis. Preferably, substrings used in the landmark finder should occur in the top right-hand corner of the figure.

Subsets of substrings were chosen by selecting a precision threshold that the substring has to surpass, as well as a threshold for the number of times a substring has to occur in PDB. Six subsets were generated for alpha helix, ten for alpha helix 3-10, eight for alpha helix pi, six for alpha helix combined and seven for extended strand. The subsets of substrings were evaluated taking into account the following four criteria: 1) average precision of all substrings in the subset, 2) the proportion of sequences in PDB matched by at least one substring, 3) the proportion of known secondary structure strings in PDB of a given secondary structure element matched at least once by a substring, 4) the correlation between the light matter scores calculated using the perfect finders described in chapter B.1.1 and the finders using the subsets of substrings. The correlations were calculated separately for the 4MTP, 2HLA, Polymerase, and HA test datasets described in chapter B.1.2. Figure B.5 shows the evaluation of the subsets of substrings according to the four criteria, for each secondary structure element. The x-axes on all plots indicate the name of the subsets of substrings under evaluation, where the first number in the name corresponds to the threshold of the number of occurrences of the substrings in PDB and the second number corresponds to the threshold for the precision of the substrings in the subset. The data for the alpha helix 3-10 and alpha helix pi are most likely confounded by the small number of those secondary structure elements in the test datasets (Table B.5). The subsets of substrings selected for use in the finder for each secondary structure element are: alpha helix: 20-0.85, alpha helix 3-10: 1-0.5, alpha helix pi: 1-0.5, alpha helix combined: 20-0.85, and extended strand: 10-0.5. I selected these subsets of substrings because they have the best combination of correlation between the light matter scores from the perfect finders and scores using the finders from the subsets of substrings across the four test datasets (Figs. B.5c, f, i, l, o), as well as the highest possible total precision (Figs. B.5a, d, g, j, m) and fraction of sequences and structures matched in PDB (Figs. B.5b, e, h, k, n). The finders that identify the substrings are called ‘Aho-Corasick’ (AC) finders, after the matching algorithm published in 1975 [410]. We refer to the features themselves as ‘AC AlphaHelix’, ‘AC AlphaHelix_3_10’, ‘AC AlphaHelix_pi’, ‘AC ExtendedStrand’, and ‘AC AlphaHelix_combined’. *The landmark finder that identifies the features using the subsets of substrings developed above was implemented by Terry Jones.*

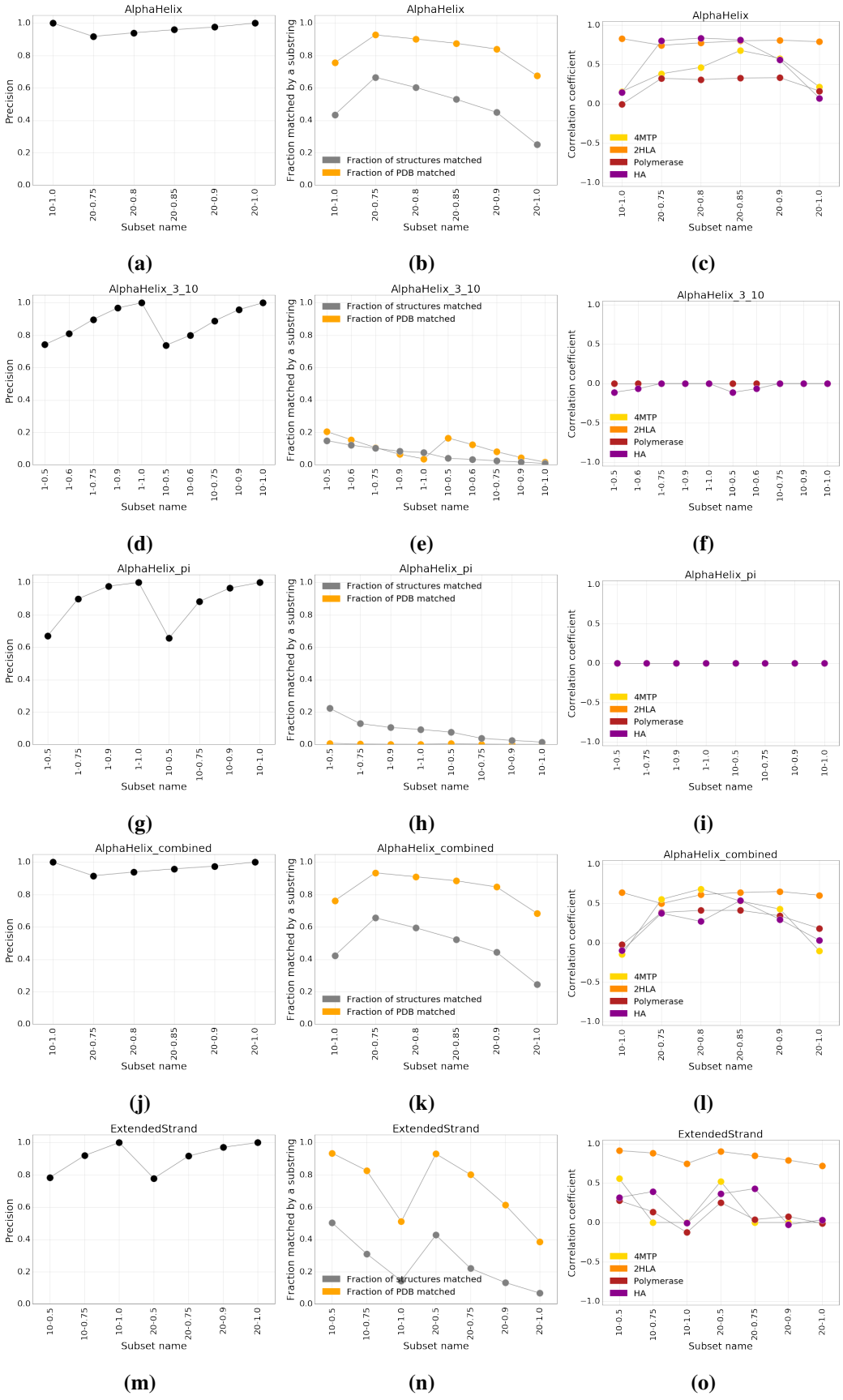

AlphaHelix
Precision
Subset name
(a)
AlphaHelix
Fraction matched by a substring
Fraction of structures matched
Fraction of PDB matched
Subset name
(b)
AlphaHelix
Correlation coefficient
4MTP
2HLA
Polymerase
HA
Subset name
(c)
AlphaHelix_3_10
Precision
Subset name
(d)
AlphaHelix_3_10
Fraction matched by a substring
Fraction of structures matched
Fraction of PDB matched
Subset name
(e)
AlphaHelix_3_10
Correlation coefficient
4MTP
2HLA
Polymerase
HA
Subset name
(f)
AlphaHelix_pi
Precision
Subset name
(g)
AlphaHelix_pi
Fraction matched by a substring
Fraction of structures matched
Fraction of PDB matched
Subset name
(h)
AlphaHelix_pi
Correlation coefficient
4MTP
2HLA
Polymerase
HA
Subset name
(i)
AlphaHelix_combined
Precision
Subset name
(j)
AlphaHelix_combined
Fraction matched by a substring
Fraction of structures matched
Fraction of PDB matched
Subset name
(k)
AlphaHelix_combined
Correlation coefficient
4MTP
2HLA
Polymerase
HA
Subset name
(l)
ExtendedStrand
Precision
Subset name
(m)
ExtendedStrand
Fraction matched by a substring
Fraction of structures matched
Fraction of PDB matched
Subset name
(n)
ExtendedStrand
Correlation coefficient
4MTP
2HLA
Polymerase
HA
Subset name
(o)

Figure B.5: Evaluation of subsets of substrings. Plots in each row show the four metrics for evaluation for each secondary structure element. Subsets of substrings evaluated are plotted on the x-axis. The first number in the subset name corresponds to the threshold of the number of occurrences and the second number corresponds to the threshold for the precision. The leftmost column of plots (a, d, g, j, m) shows the total precision of a subset, calculated by averaging out the precision of all substrings. The middle column of plots (b, e, h, k, n) shows the fraction of sequences in PDB that was matched by at least one substring in orange, and the fraction of secondary structures of a given secondary structure element in perfect matched by at least one substring in grey. The rightmost column (c, f, i, l, o) shows the correlation coefficients between scores obtained using the PDB finders and scores obtained using the substrings of the respective subsets. Correlation coefficients are calculated separately for the four test datasets.

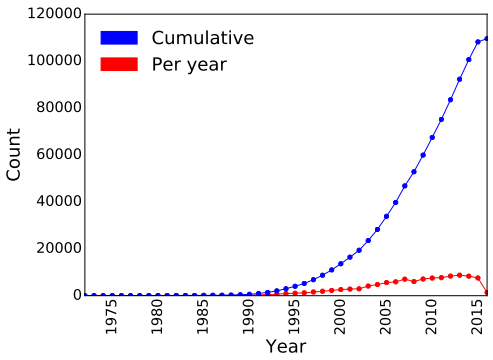


Figure B.6: Structures added to PDB between 1972 and 2016. The blue line indicates the total number of structures present in PDB at a particular time. The red line indicates the number of structures that were added in that particular year.

PDB almost certainly is an incomplete representation of all protein structures on earth and it can be expected to grow in the future. This is illustrated in Fig. B.6, which shows the cumulative number of structures added to PDB in blue and the number of structures added per year in red. It is apparent that the growth of PDB cannot be expected to slow down in the near future. PDB is therefore highly incomplete, and it is important to investigate how the subsets of substrings selected to be used in the Aho-Corasick finders generalise to sequences not currently in PDB. I used the sequences present in PDB up to each year between 1972 and 2016 (for example 1972–1973, 1972–1974 etc.), and selected subsets of substrings using same parameters as used for the subsets that were found to perform best according to the evaluation criteria, resulting in 44 subsets per secondary structure element, one for each year. I evaluated each subset of substrings for its performance on sequences that were added to PDB in future years. So, the subset generated using the sequences present in PDB between 1972 and 1980 was evaluated against the sequences present in PDB between 1972 and 1980, 1972 and 1981 and so on. I calculated the quotient of correctly identified secondary structures by the total number of secondary structures present in PDB at the time for each year (referred to as the 'fraction of known structures'). The results for each secondary structure element are shown in Fig. B.7. Empty circles show the fraction of known structures present in PDB in the year up to which the subset of substrings was selected. The blue line extending from each empty circle indicates how this subset of substrings performs into the future, as measured by the fraction of known structures present in PDB in future years. The coloured line indicates how the subset of substrings selected up to 2016 performs against structures in PDB selected in each year into the past. The alpha helix, alpha helix combined and extended strand show a decreased increase in performance after 1994 and 1998 respectively, suggesting that the current set may be an accurate representation of the existing but undiscovered secondary structures. For the alpha helix 3-10 and alpha helix pi secondary structures, no drop off is visible, probably influenced by the low number of these secondary structure elements (Table B.5). This suggests that the

subsets of substrings for these two secondary structure elements does not yet contain a representative set of all secondary structure substrings.

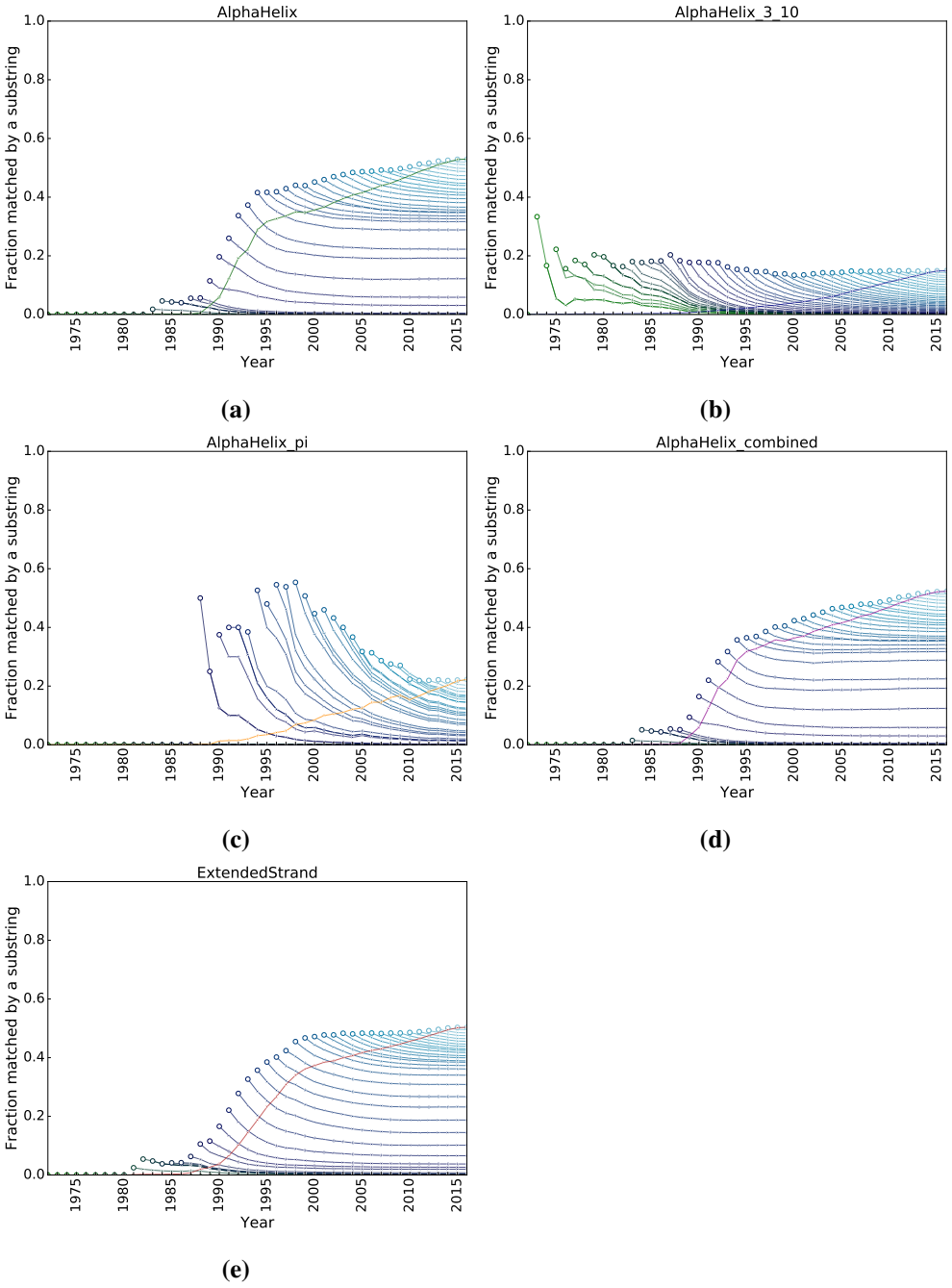
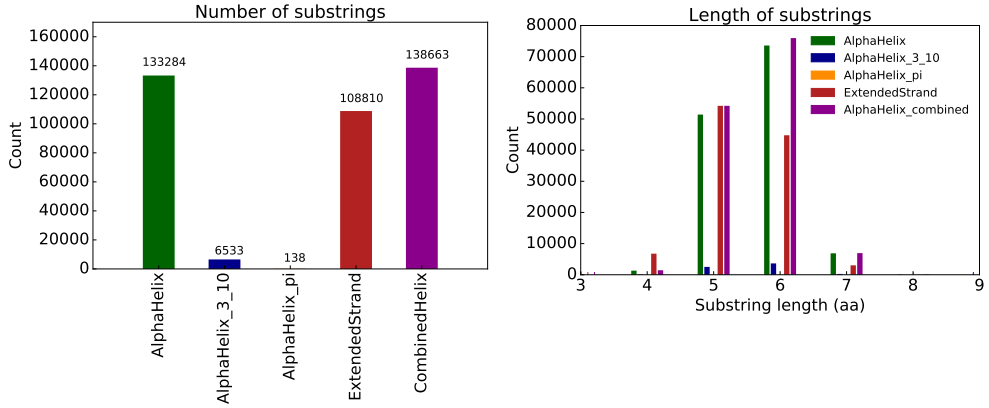


Figure B.7: Performance of subsets of substrings selected from all sequences present in PDB from each year using the criteria that were determined on the set from 2016. Performance was determined using the subset of substrings from each particular year on the subset of sequences present in PDB in all future years. Performance was measured by the fraction of known structures a given subset of substrings can identify. Empty circles show the fraction of known structures present in PDB in the year up to which the subset of substrings was selected. The blue line extending from each empty circle indicates how this subset of substrings would perform into the future, as measured by the fraction of known structures present in PDB in future years. The coloured line indicates how the subset of substrings selected up to 2016 performs against structures in PDB selected in each year into the past.

	AC AlphaHelix	AC AlphaHelix_3_10	AC AlphaHelix_pi	AC AlphaHelix_combined	AC ExtendedStrand
AC AlphaHelix	-				
AC AlphaHelix_3_10	0	-			
AC AlphaHelix_pi	0	0	-		
AC ExtendedStrand	0	0	0	-	
AC AlphaHelix_combined	132,663	377	5	0	-

Table B.6: Overlap of substrings in the subsets used in the Aho-Corasick finders.



(a) **(b)**

Figure B.8: Number and length of substrings in the final subsets, by secondary structure element. a) The number of substrings in the final subsets, per secondary structure element. b) The length of substrings in the final subsets, per secondary structure element. For all secondary structure elements, the lengths fall in a range from four to eight amino acids, with peaks at five or six amino acids.

Figure B.8a shows the number of substrings present in each subset and Fig. B.8b shows the length distribution within each subset. The majority of substrings are five or six amino acids long. The AlphaHelix and AlphaHelix_combined subsets both contain a similar number of substrings of similar lengths. Figure B.8a also suggests that the finders may be of limited use for the identification of AlphaHelix_3_10 and AlphaHelix_pi, due to the small number of substrings representing those secondary structure element in the subset. Table B.6 shows the number of substrings that are in common between the five subsets of substrings. Not surprisingly, all individual alpha helix subsets have some overlap with the AlphaHelix_combined subset. The AC_AlphaHelix_combined and the other three AC_AlphaHelix finders should not be used together, since it would result in the same features being identified twice. There is no overlap between the subsets of the other secondary structure elements.

The new finders were evaluated using the test datasets described in chapter B.1.2. All matches were evaluated using the perfect finders described in chapter B.1.1 and the Aho-Corasick finders described above. Figure B.9 shows a scatter plot of the scores calculated using the Aho-Corasick finders with the subsets described above, and the perfect finders. The blue line and blue shaded area indicate the linear regression and 95% confidence interval. The figure shows that there is a correlation between scores obtained using the perfect finders and the scores obtained from using the new Aho-Corasick finders, suggesting that the Aho-Corasick finders identify adequate features for sequence comparison in the light matter algorithm.

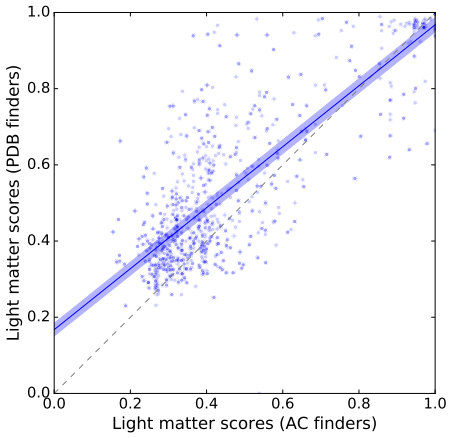


Figure B.9: Correlation between the scores obtained from using the perfect PDB finders and the scores obtained using the Aho-Corasick finders. Scores were calculated using the test datasets described in chapter B.1.2. For visual reference, the dashed line indicates perfect correlation. The blue line shows the linear regression and 95% confidence interval. Slope and correlation coefficient are: 0.8 and 0.83 (R^2=0.68). Perfect finders: 'PDB AlphaHelix', 'PDB AlphaHelix_3_10', 'PDB AlphaHelix_pi', and 'PDB ExtendedStrand'. Aho-corasick finders used: 'AC AlphaHelix', 'AC AlphaHelix_3_10', 'AC AlphaHelix_pi', and 'AC ExtendedStrand'.

B.2.1.2 Landmarks based on sequence motifs

A sequence motif is a short pattern in the nucleotide or amino acid sequence that is thought to have a particular biological function [411]. The same motifs often occur in proteins with functional and structural similarity [411]. The view of protein structures as being rigid and stable is increasingly being recognised as flawed [412]. Up to 50% of amino acids in eukaryotic sequences are predicted to be disordered [412], meaning that they do not adopt a fixed three-dimensional structure. Further, in viral proteomes, between 15 – 23% of segments <30 amino acids in length are disordered [413]. Some sequence motifs occur predominately in disordered regions of proteins [414]. Given that motifs are often associated with similar structure and function, and may help identifying reads matching against disordered regions of a protein, I implemented

two landmark finders. One is based on the PROSITE [349] and the other on the Eukaryotic Linear Motif [352] database.

Landmarks based on PROSITE motifs

The PROSITE database contains entries that describe protein domains, families, and functional sites. For each of these entries, the database also contains the associated patterns and profiles to identify them [349]. Patterns are between 10 to 20 amino acids long and include enzyme catalytic sites, prostethic group attachment sites, amino acids involved in binding a metal ion, cysteines involved in disulfide bonds, and regions involved in binding a molecule or another protein [415]. Patterns are described by regular expressions. Profiles have a numerical weight for each possible match or mismatch between a sequence residue and a profile position. Profiles have an advantage over patterns since they allow for mismatches [415], but they characterise protein domains over their entire length. This makes them unsuitable as landmarks, and I focus on patterns instead. The PROSITE database used in the light matter algorithm (downloaded on 11 November 2015) contains 1,309 patterns.

Landmarks based on short linear motifs

Short linear motifs (SLiMs) are three to eleven amino acids long and are generally situated in disordered regions of a protein. They mediate protein-protein interactions, broadly acting either as ligand binding sites or modification sites [414]. Their short length and location in disordered regions of proteins make SLiMs interesting features to use as landmarks. First, because their short length makes it possible to identify them in NGS reads, and second, because they allow me to identify features in disordered regions of proteins, in contrast to the other finders described so far. The eukaryotic linear motif (ELM) database is a manually curated database of SLiMs [352]. Each motif is described by a regular expression pattern. As of 13 July 2016, the ELM database contains 251 patterns. 49 of those are specific for viral proteins. I wrote a finder that detects SLiMs using the regular expression patterns in the ELM database. The finder can use either the complete set of 251 patterns, or the 49 patterns that are specific to viruses. The finder was evaluated using the 2HLA, 4MTP, Polymerase, and HA test datasets described in chapter B.1.2. Figure B.10 shows that using just the virus-specific patterns leads to less matches being assigned a light matter score of 0.0, and a better correlation between the scores computed by the light matter algorithm and the Z-scores inferred by the Dali algorithm. Hence, I only use the virus specific database in the feature finder.

There is only very limited overlap between the PROSITE and the ELM databases, only one pattern (‘RGD’) is present in both.

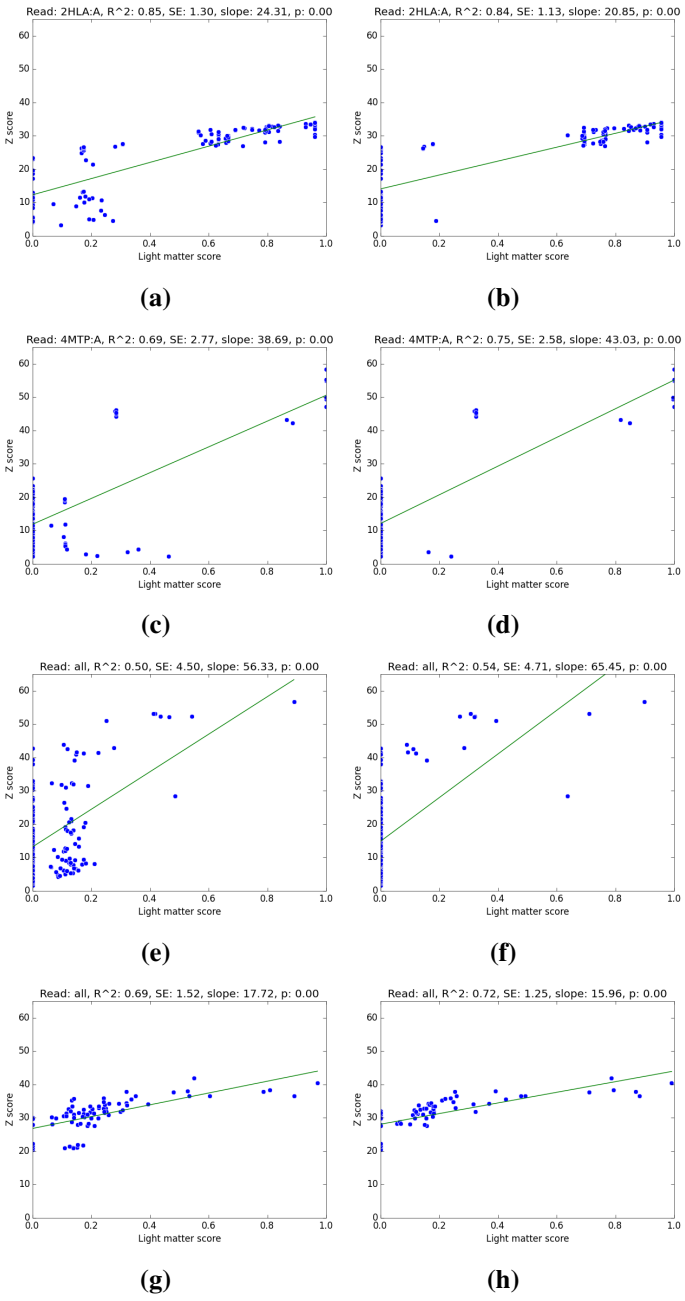


Figure B.10: Evaluation of the different ELM subsets. The left column shows the evaluations using the viral ELMs only and the right column shows all ELMs. Each row is one of the evaluation datasets described in chapter B.10. Figures a) and b) are the 2HLA dataset, Figs. c) and d) the 4MTP dataset, Figs. e) and f) the Polymerase dataset and Figs. g) and h) the HA dataset. The y-axes on all plots shows the Dali Z-scores, and the x-axes indicate the light matter score calculated by the algorithm.

B.2.1.3 Landmarks based on amino acids

Some amino acids are more conserved than others. Scoring matrices represent the relative rates of substitutions, showing how likely one amino acid is to be replaced by another. Two different kinds of scoring matrices are widely used, the PAM (Point Accepted Mutations) [350] and BLOSUM (BLOcks SUbstitution Matrix) [351, 394] matrices. The PAM1 matrix was constructed from proteins that were 85% or more similar to one another. To simulate different degrees of divergence, the PAM matrix can be multiplied by itself [394]. Higher numbers in the PAM matrix name refer to larger evolutionary distance. BLOSUM matrices were constructed from a database of aligned proteins. For the BLOSUM62 matrix, only protein sequences with 62% or more sequence conservation were used to calculate the frequencies in the scoring matrix. The larger the suffix in the BLOSUM matrix name, the shorter the evolutionary distance [416]. Both matrices agree that tryptophan and cysteine are the most conserved amino acids (see Fig. B.11). Tryptophan has a large aromatic side chain and cysteine participates in the formation of disulfide bonds. I therefore decided to use cysteines as landmarks and tryptophans as a trig points in the light matter algorithm. It is possible to use one or multiple additional amino acids as landmarks or trig points, however I did not test those combinations. One has to keep in mind that the more different amino acids are used in finders, the more the algorithm will behave as if the sequence comparisons are done at the amino acid level.

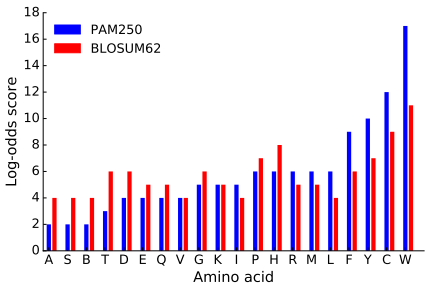


Figure B.11: Log-odds scores for each amino acid in the PAM250 and BLOSUM62 matrices. Scores were taken from Dayhoff *et al.* [350] and Henikoff *et al.* [351]. Amino acids are sorted by their log-odds score in the PAM250 matrix, and shown on the x-axis by their one-letter abbreviation.

B.2.2 Trigonometry points

Trigonometry (trig) points are the second group of features used by the light matter algorithm, alongside landmarks. A trig point can only form a pair with a landmark, not with another trig point. The trig points currently implemented in the algorithm are either based on amino acids or on specific amino acid property patterns (see table B.3).

B.2.2.1 Trig points based on amino acids

This trig point finder is analogous to the landmark finder based on amino acids described in chapter B.2.1.3. The tryptophans in a sequence are used as trig points, as this amino acid is highly conserved (see Fig. B.11).

B.2.2.2 Trig points based on amino acid properties

Each amino acid has specific biochemical properties, which may influence protein evolution. If an amino acid is replaced in a context where the structure needs to be conserved, the replacing amino acid often has similar properties [417]. Based on this idea, I implemented two types of trig point finders. The 'Peaks' and 'Troughs' and the 'IndividualPeaks' and 'IndividualTroughs' finders. Peaks and Troughs are identified by sliding a window of length three along the amino acid sequence. If the sum of the property values for multiple predetermined amino acid properties is higher / lower for the current amino acid than for the two amino acids on either side, the current amino acid is a Peak / Trough. Similarly, IndividualPeaks and IndividualTroughs are identified by sliding a window of length three along the amino acid sequence. If the property values for multiple predetermined amino acid properties are all higher / lower for the current amino acid than for the two amino acids on either side, the current amino acid is an IndividualPeak / IndividualTroughs.

Xia *et al.* (1998) studied mitochondrial genes and found that the genetic code appears to have evolved towards minimising polarity and hydropathy. They also found that proteins encoded by mitochondrial DNA have more amino acids with typical composition and isoelectric point values, so that non-synonymous mutations result in small differences in those properties [417]. Based on their work, I decided to use chemical composition, isoelectric point, and polarity in the trig point finders.

To investigate the behaviour of the Peak, Trough, IndividualPeak, and IndividualTrough trig point finders, I used them on sequences from the following virus groups:

49 bunya-, arena-, and orthomyxovirus sequences, 304 amino acids in length, 65 influenza haemagglutinin sequences, 549 amino acids in length, 94 *Mononegavirales* sequences, 945 amino acids in length, 29 *Nidovirales* sequences, 210 amino acids in length, and 23 rhabdovirus sequences, 350 amino acids in length. Sequences other than the influenza haemagglutinin sequences were provided by Christian Drosten. The influenza haemagglutinin sequences were provided by David Burke. I was interested to investigate whether specific amino acids would preferentially be identified as trig points. Figure B.12 shows the frequencies at which a particular amino acid is found in a Peak, Trough, IndividualPeak, IndividualTrough, in both a Peak and IndividualPeak, Peak and IndividualTrough, Trough and IndividualPeak, or Trough and IndividualTrough, or is not used in a trig point.

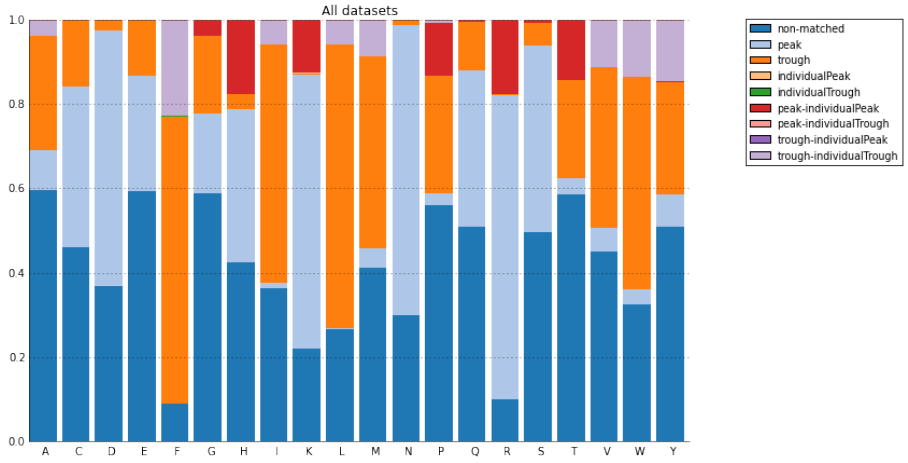


Figure B.12: Frequencies at which particular amino acids are found by a particular trig point finder. Frequencies are shown on the y-axis, and the different amino acids on the x-axis, denoted by the one-letter amino acid abbreviations.

The following observations can be made. An IndividualPeak is almost always also a Peak and and IndividualTrough is almost always also a Trough. The same is not true for Peaks and Troughs. Some amino acids are more frequently assigned to a Peak or Trough: arginine (R) is most commonly found in Peaks and IndividualPeaks, together with lysine (K), asparagine (N), and aspartic acid (D); phenylalanine (F) is most commonly found in Troughs and IndividualTroughs, together with leucine (L), tryptophan (W), and isoleucine (I). Some amino acids do not form a trig point around 60% of the time, including alanine (A), glutamic acid (E), glycine (G), proline (P), and threonine (T). Some amino acids can be found in both Peaks and Troughs, including cysteine (C), glutamic acid, and glutamine (Q), however, those are also the amino acids that often are not part of a trig point at all.

Given the overlap between Peaks and IndividualPeaks and Troughs and IndividualTroughs, IndividualPeaks and IndividualTroughs do not seem add information when comparing two sequences other than that provided by the Peaks and Troughs and do not need to be used by the algorithm.

B.3 METHODS TO DETERMINE THE SIGNIFICANCE OF A MATCH

After all features have been identified and paired, the pairs of the subject and the query are compared. If a pair exists in both the query and the subject, the offsets of the locations of the first feature in the pair in the query and the subject are subtracted, and these 'deltas' are entered in a histogram. Subsequently, the histogram is assessed to decide whether the comparison between the query and the subject can be considered a significant match or not. A significant match should have a distinctive peak in its histogram, as this indicates a high number of identical pairs with the same relative sequence offsets in the query and the subject. Note that 'significant' here is used informally, and does not imply formal statistical significance. I implemented three different methods for assessing whether a histogram has one or multiple bins that are indicative of a significant match, so-called 'significance methods'. All significance methods consider each histogram bin separately and assess whether it is significant. If no significant bin is present in a histogram, the match as a whole is insignificant and is not considered further. If a significant bin is found, a score is calculated (see chapter B.4). Each significance method calculates a 'significance cut-off', corresponding to the minimum bin height in order for a bin to be significant. The term 'bin height' refers to the number of pairs or amino acids in pairs in a histogram bin. Figure B.13 illustrates the effects of using the different significance methods on the comparison between two *Bunyaviridae* polymerase sequences. Significant bins are shown in red, insignificant bins in black. The three significance methods are:

MaxBinHeight: the significance cut-off value is the number of pairs in the second-highest bin when the query is compared against itself. In the histogram from the comparison of the query and the subject, every bin higher than this cut-off value is considered significant (Fig. B.13a).

HashFraction: the significance cut-off is calculated based on the theoretical maximum number of pairs that could be present in the highest bin if there was a perfect match between query and subject, multiplied by a user-defined 'significance fraction' between 0.0 and 1.0. The theoretical maximum number of pairs that can be present in the highest bin is equal to the total number of pairs in either the subject or the query, depending on which of the two has the smaller number of pairs. Any bin that is higher than the significance cut-off is considered significant (Fig. B.13b).

AAFraction: the AAFraction significance method uses the same principle as the HashFraction significance method, but rather than considering the number of pairs in

the bin, it considers the number of amino acids in pairs in the bin. The significance cut-off is calculated by using the theoretical maximum number of amino acids in pairs that can be present in the highest bin in a perfect match, multiplied by a user-defined 'significance fraction' between 0.0 and 1.0. Note that each amino acid is only counted once, even if it occurs in multiple features that form pairs in a bin (Fig. B.13c).

In the following sections I describe the influence of the significance methods and the cut-offs employed on the number of matches that are identified as significant by the significance methods. I also describe the difference between the AAFraction and the HashFraction significance methods.

B.3.1 Matches considered significant by different significance methods

The ideal significance method should identify matches between sequences coding for proteins with similar structures as significant, and it should disregard matches between sequences coding for proteins with dissimilar structures. As a general rule, a Dali Z-score above 20 indicates the two structures are definitely homologous, between 8 and 20 means the two are probably homologous, between 2 and 8 is a grey area, and a Z-score below 2 is not significant [359]. For our purposes, any significance method and significance cut-off combination that finds a match with Dali Z-scores above 20 and between 8 and 20 to be insignificant is not desirable. In order to investigate whether the significance methods are effective at marking matches with Z-scores <8 as insignificant, while keeping matches with higher Z-scores, I performed the following experiment. I used all pairwise comparisons in the 2HLA, 4MTP, and 4PH0 test datasets to evaluate which of them are considered significant by different combinations of significance method and significance cut-off. For each pairwise comparison, I determined whether the match is considered significant by the MaxBinHeight, HashFraction, and AAFraction significance methods.

Figures B.14b–d) show the fraction of matches that are considered insignificant using the MaxBinHeight, HashFraction, and AAFraction significance methods, and, in case of the HashFraction and AAFraction methods, also show the effects of different significance cut-off values. For reference, Fig. B.14a shows the number of matches in the test dataset with a particular Z-score as black vertical bars.

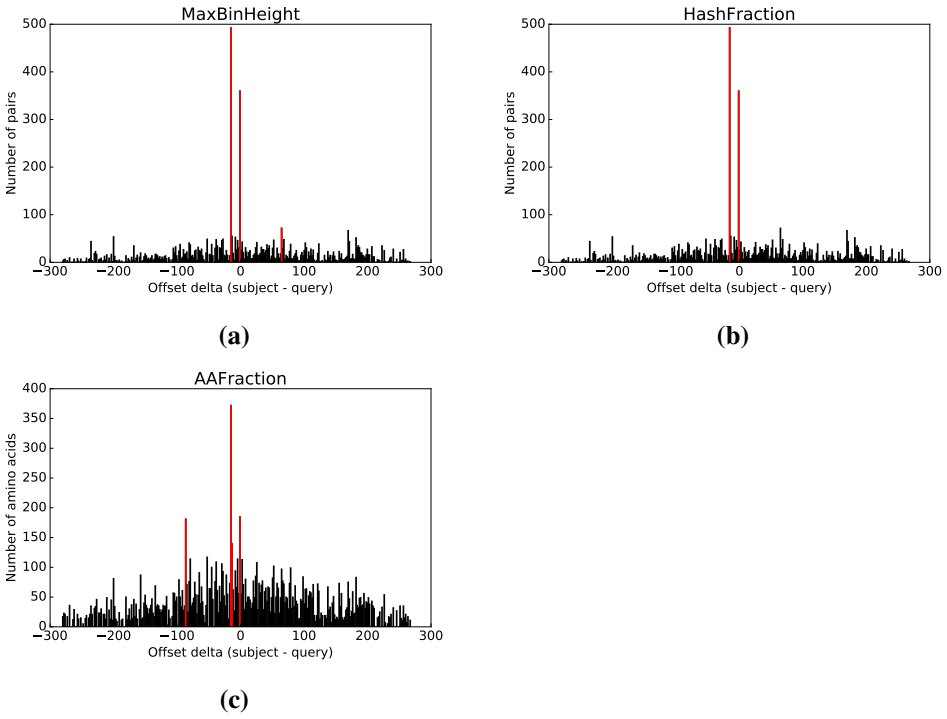


Figure B.13: Illustration of the effect of using different significance methods on the comparison between two *Bunyaviridae* (bunyamveravirus and oropuchevirus) polymerase sequences. Significant bins are shown in red, insignificant bins in black. The sequences have an amino acid sequence similarity of 67%. **a) MaxBinHeight. b) HashFraction.** The significance fraction was set to 0.03. **c) AAFraction.** The significance fraction was set to 0.2. Note that because the AAFraction significance method calculates the height of the histogram bins by the number of amino acids rather than the number of pairs in the bin, the height of the histogram bins in Fig. c) differs from figures a) and b).

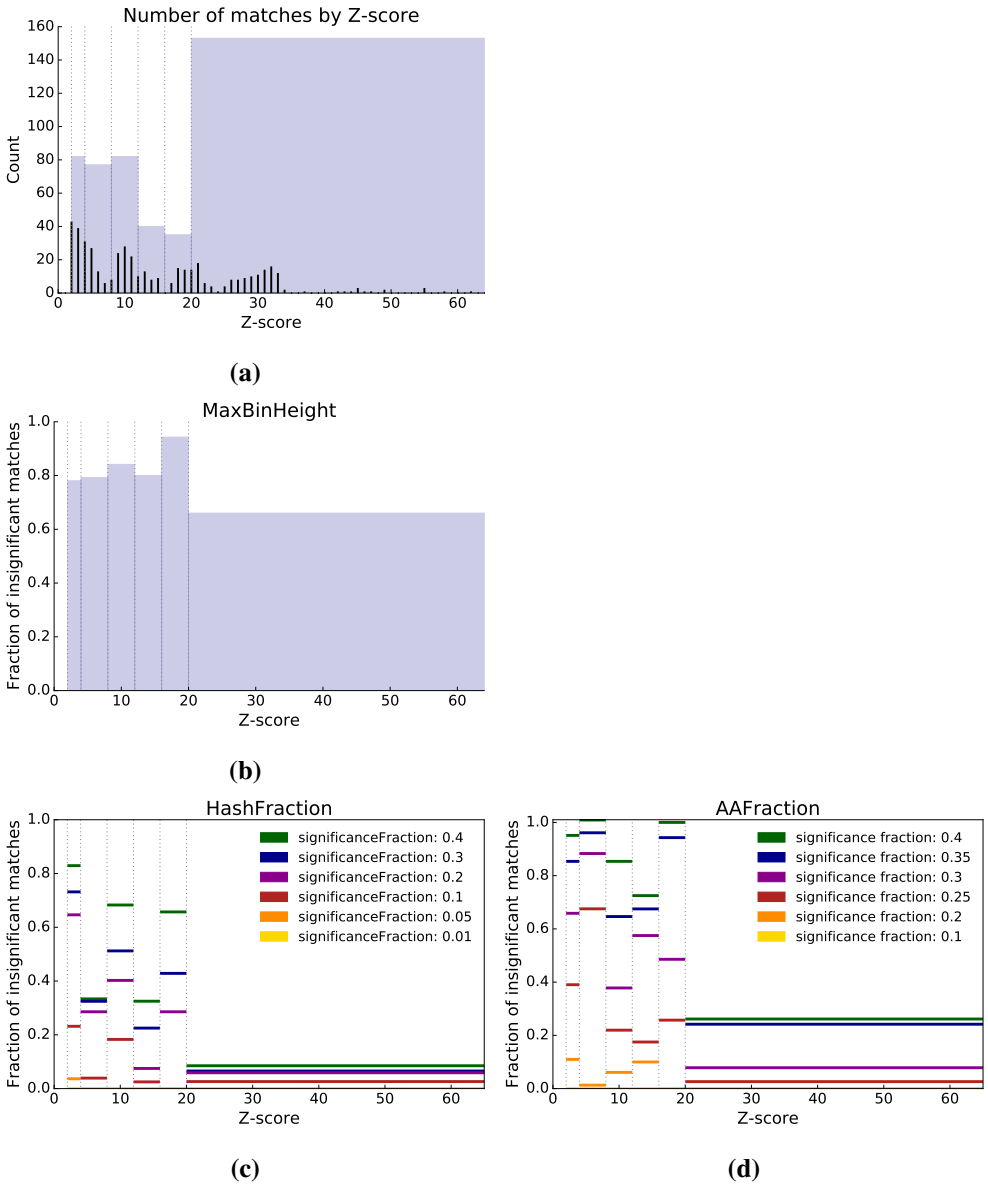


Figure B.14: The fraction of matches considered insignificant using different significance methods. To calculate the match, the 'AC AlphaHelix_combined', 'AC ExtendedStrand' finders and no trig point finders were used. Z-scores were binned from 0–1, 2–3, 4–7, 8–11, 12–15, 16–19, and >20 indicated by grey dotted lines. **a) The total number of matches in the test datasets with particular Z-scores are shown as black vertical bars.** Blue rectangles show the number of Z-scores in the range of 0–1, 2–3, 4–7, 8–11, 12–15, 16–19, and >20. **b) The fraction of insignificant matches using the MaxBinHeight significance method. c) The fraction of insignificant matches using the HashFraction significance method. d) The fraction of insignificant matches using the AAFraction significance method.** In figures b)–d), the y-axis indicates the fraction of insignificant matches, out of all matches in a particular Z-score bin. In figures c) and d), different colours show the results using different significance fractions.

For a less fine-grained representation, I binned the Z-scores from 0–1, 2–3, 4–7, 8–11, 12–15, 16–19, and >20, denoted by dotted vertical lines or blue shaded areas. The MaxBinHeight significance method excludes at least 66% of all matches, even those with Z-scores >20 (Fig. B.14b). This behaviour is undesirable, and this significance method was not used in the further development of the algorithm. The HashFraction and AAFraction significance methods omit between 0.0 to 66%, and 0 to 94% of the matches with Z-scores between 16–19, respectively, depending on the significance fraction (Figs. B.14c–d). Considering the effect of the significance fraction, unsurprisingly, the higher the significance fraction, the more comparisons are marked insignificant. Further, it appears that using a significance fraction of 0.2 for the HashFraction significance method and a significance fraction of 0.25 for the AAFraction significance method results in the desired effect of ruling out a large proportion of the bad and grey-area comparisons with Z-scores between 0 and 8, while including most of the matches with Z-scores above 20. The fractions of comparisons marked insignificant does not decrease linearly with increasing Z-score. For example, the HashFraction significance method with a significance fraction of 0.4 only marks about 30% of all comparisons with Z-scores between 4–7 as insignificant, whereas 68% of all comparisons with Z-scores between 8–11 are marked as insignificant. When considering each of the three test datasets separately (Fig. B.15) using the HashFraction significance method with a significance fraction of 0.4, it becomes apparent that there is a large variation in the proportion of matches that are excluded in each Z-score bin between datasets. For example in the Z-score bin between 16–19, 63% of all matches in the 4MTP dataset are insignificant, while in the 2HLA dataset, almost all matches are significant. What causes this behaviour is currently unknown, and could be investigated further. Therefore, there is no indication that a single combination of HashFraction or AAFraction significance methods and significance cut-offs will perform best under all circumstances. This has to be kept in mind when setting significance fractions and analysing results.

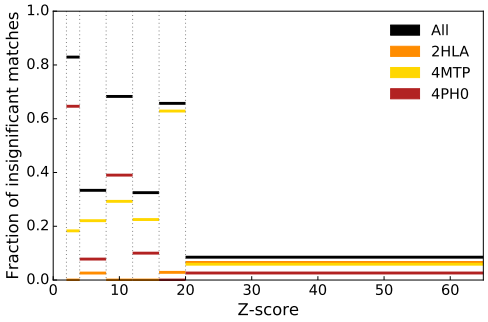


Figure B.15: The fraction of matches considered insignificant using the HashFraction significance method, for different test datasets. The significance fraction was set to 0.4.

B.3.2 HashFraction versus AAFraction histograms

The HashFraction and AAFraction significance methods differ in the way their histogram bin heights are calculated. The HashFraction significance method determines the bin height by counting the number of pairs in each bin, while the AAFraction significance method uses the number of amino acids in pairs in each bin. Each amino acid is counted only once, even if it occurs in multiple pairs in the same bin. Figure B.16 shows the effect of these two ways of calculating the bin height on the shape of the histogram. Three matches were performed with sequences of different relatedness. Figures B.16a–b show a bunyamvera (BUNV) sequence compared against itself, Figs. B.16c–d show the BUNV sequence compared to an oropuchevirus (OROV) sequence (the amino acid sequences have an identity of 67%), and Figs. B.16e–f show the BUNV sequence compared to a kibalevirus (KIBV) sequence (the amino acid sequences have an identity of 45%). Figures B.16a, c, and e) show the HashFraction histograms, and Figs. B.16b, d, and f) the AAFraction histograms. The AC AlphaHelix_combined, AC ExtendedStrand finders and no trig point finders were used. When comparing the HashFraction and AAFraction histograms, it is apparent that the histogram peaks are more pronounced in the HashFraction histograms. This can be attributed to the fact that all pairs are counted for the HashFraction histogram bin height, even if they cover the same amino acids. This is for example the case when the same landmark forms pairs with multiple trig points. Furthermore, each pair is weighted equally, no matter how many amino acids it contains. In the AAFraction significance method, on the other hand, each amino acid is counted once to determine the bin height, even if it is part of multiple pairs, thus flattening out the peak. The bin height in the AAFraction significance method is biased towards bins that contain pairs with features spanning many amino acids. So, a pair that consists of an amino acid landmark and an amino acid trig point is weighted less than a pair that consists of an ‘AC alphaHelix’ landmark (which always contains at least four amino acids) and an amino acid trig point. This behaviour is offset to some extent by the fact that each amino acid is only counted once. The HashFraction significance method on the other hand is agnostic to the type and size of the features in the pairs it contains. Given that the HashFraction leads to more obvious histogram peaks, using it as the default significance method is more desirable.

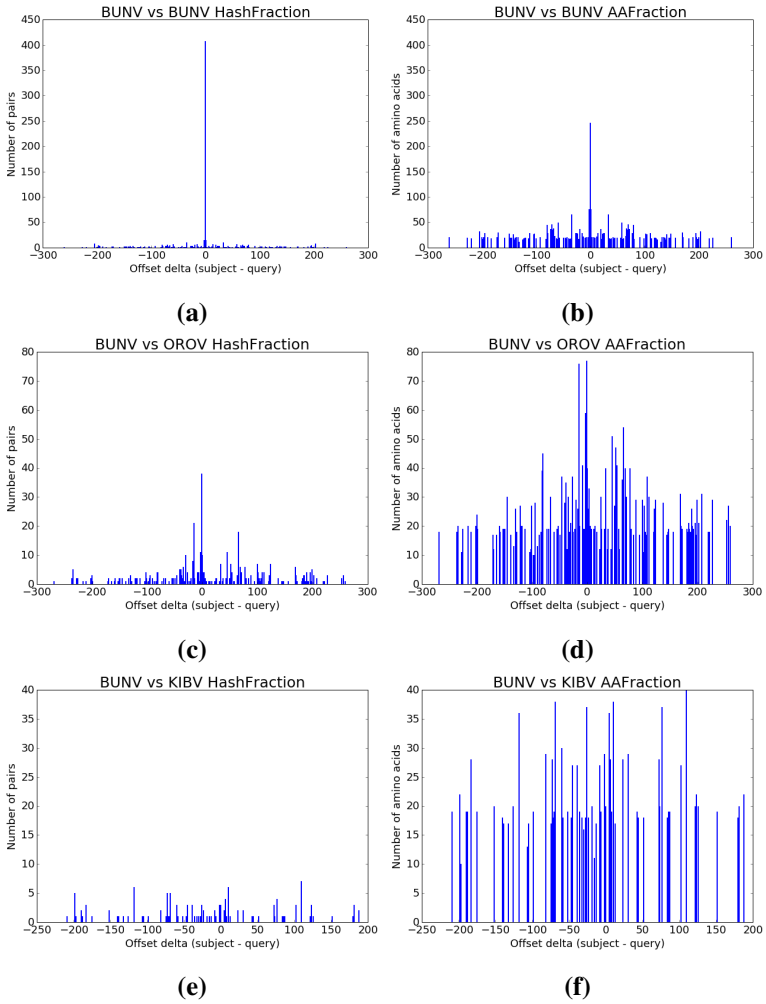


Figure B.16: Histograms for the HashFraction and AAFraction significance methods. Three matches were performed with sequences of different relatedness. **Figures a–b) show a bunyamvera virus (BUNV) sequence compared against itself, Figs. c–d) show the BUNV sequence compared to an oropuchevirus (OROV) sequence, and Figs. e–f) show the BUNV sequence compared to a kibalevirus (KIBV) sequence.** Figures a), c) and e) show the HashFraction histograms, Figs b), d), and f) the AAFraction histograms. The 'AC AlphaHelix_combined', 'AC ExtendedStrand' finders and no trig point finders were used.

B.4 METHODS FOR SCORING MATCHES

Assigning a meaningful score to a match is essential for the interpretation of the results generated by the light matter algorithm. Given that we are comparing sequences using predicted structural features, it is desirable that the score should correlate with conventional measures of structural similarity, under the assumption that these measures correlate with the structural reality.

The significance methods presented in chapter B.3 report which histogram bins from a match are significant. For those bins, a score is calculated. Two types of scores can be calculated: firstly a so-called 'bin score', in which a score is calculated for each significant bin of a match, and only the highest score is reported as the score of the match, and secondly, an 'overall score', which incorporates some or all bins that are considered significant into the score calculation. All scoring methods assign a score between 0.0 and 1.0, with 0.0 being the worst, and 1.0 being the best score.

B.4.1 Bin scoring methods

In the bin scoring methods, a score is calculated individually for all significant bins of a match and the highest score is used as the score to reflect the similarity of the two sequences that are compared. I implemented two bin scoring methods in the light matter algorithm:

The **MinHashesScore** scoring method is based on the idea that in a perfect match, the maximum number of pairs in the best bin is the number of pairs found in the subject or the query, depending on which of the two has fewer pairs. This case should receive the best score of 1.0, and worse matches should be assigned a score lower than that. Consider the schematic match in Fig. B.17: grey dotted lines denote amino acid sequences of the query and the subject. The grey shaded area is what I call the 'matched region'. It is the area of the sequence that contains the features that are involved in the match between the query and the subject, shown as black boxes, as well as features that are not involved in the match, shown as orange boxes. Light green boxes represent unmatched features outside the matched region. Trig points are omitted for clarity. The MinHashesScore is calculated by dividing the number of matching pairs in the bin by the total number of matching pairs in either the query or subject, depending on which contains less pairs (equation B.2).

$$MinHashesScore = \frac{M_1 + M_2}{min(U_{S1} + M_{S1} + B_{S1} + M_{S2} + U_{S2} + U_{S3} + U_{S4}, U_{Q1} + M_{Q1} + B_{Q1} + M_{Q2} + U_{Q2})} \quad \text{(B.2)}$$

For the example in Fig. B.17, the MinHashesScore is calculated as:

$$MinHashesScore = \frac{2}{min(5, 7)} = 0.4 \quad \text{(B.3)}$$

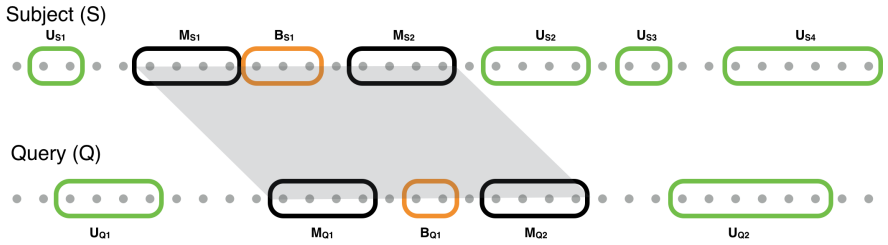


Figure B.17: Schematic of a match to illustrate the bin score calculation. Horizontal grey dotted lines denote the amino acid sequences. The grey shaded area is the matched region. Light green rectangles represent unmatched features outside the matched region, orange rectangles represent unmatched features inside the matched region. Black rectangles represent matching features inside the matched region. Trig points are omitted for clarity.

The **FeatureAAScore** consists of a product of the score for the matched region (MatchedRegionScore) and a length normaliser (LengthNormaliser) (equation B.4). The score for the matched region is a fraction, where the numerator consist of the number of all amino acids in features in the subject and the query that match and the denominator consists of the number of all amino acids in features in the subject and the query in the region of the bin, regardless of whether they match (equation B.5). The length normaliser is the count of all amino acids in features in the region of the bin divided by the count of all amino acids in features in the whole sequence, for either the query or the subject, depending on which fraction is higher (equation B.6). The length normaliser is used to give more weight to matches that span a large part of the query and / or the subject. Note that the bin with the highest score may not be the bin with the highest number of pairs.

$$FeatureAAScore = MatchedRegionScore \times LengthNormaliser \quad \text{(B.4)}$$

$$MatchedRegionScore = \frac{M_{S1} + M_{S2} + M_{Q1} + M_{Q2}}{M_{S1} + B_{S1} + M_{S2} + M_{Q1} + B_{Q1} + M_{Q2}} \tag{B.5}$$

$$LengthNormaliser = max(\frac{M_{S1} + M_{S2} + B_{S1}}{U_{S1} + M_{S1} + B_{S1} + M_{S2} + U_{S2} + U_{S3} + U_{S4}}, \frac{M_{Q1} + B_{Q1} + M_{Q2}}{U_{Q1} + M_{Q1} + B_{Q1} + M_{Q2} + U_{Q2}}) \tag{B.6}$$

For the example in Fig. B.17, the equations and score calculations are as follows:

$$FeatureAAScore = \frac{16}{21} \times max(\frac{11}{25}, \frac{10}{20}) = 0.762 \times 0.5 = 0.381 \tag{B.7}$$

B.4.2 Overall scoring methods

The **GreedySignificantBinScore** is an overall scoring method. It is calculated using the same principle as the FeatureAAScore, except that multiple bins are taken into account. Bins are added to the overall score calculation sequentially, starting with the bin with the highest FeatureAAScore, then the second highest and so on. The overall score is first set to the highest FeatureAAScore of all bins. Subsequent bins are then incorporated into the overall score, starting with the bin with the next highest FeatureAAScore, until the overall score does not increase anymore. At this point, the overall score calculation concludes. Consider the example shown in Fig. B.18, where two significant bins, indicated by two grey shaded areas, are included in the calculation of the overall score. The equations for overall score calculation are as follows:

$$GreedySignificantBinScore = MatchedRegionScore \times LengthNormaliser \tag{B.8}$$

$$MatchedRegionScore = \frac{M_{S1} + M_{S2} + M_{S3} + M_{Q1} + M_{Q2} + M_{Q3}}{M_{S1} + M_{S2} + M_{S3} + M_{Q1} + M_{Q2} + M_{Q3} + B_{S1} + B_{Q1}} \tag{B.9}$$

$$LengthNormaliser = max(\frac{M_{S1} + M_{S2} + M_{S3} + B_{S1}}{M_{S1} + M_{S2} + M_{S3} + B_{S1} + U_{S1} + U_{S2} + U_{S3}}, \frac{M_{Q1} + M_{Q2} + M_{Q3} + B_{Q1} +}{M_{Q1} + M_{Q2} + M_{Q3} + B_{Q1} + U_{Q1}}) \tag{B.10}$$

For the example in Fig. B.18, the score is calculated as follows:

$$MatchedRegionScore = \frac{4+4+6+4+4+6}{4+4+6+4+4+6+3+2} = \frac{28}{33} = 0.848 \tag{B.11}$$

$$LengthNormaliser = max(\frac{4+4+6+3}{4+4+6+3+2+4+2}, \frac{4+4+6+2}{4+4+6+2+4}) = 0.8 \tag{B.12}$$

$$GreedySignificantBinScore = 0.848 \times 0.8 = 0.678 \tag{B.13}$$

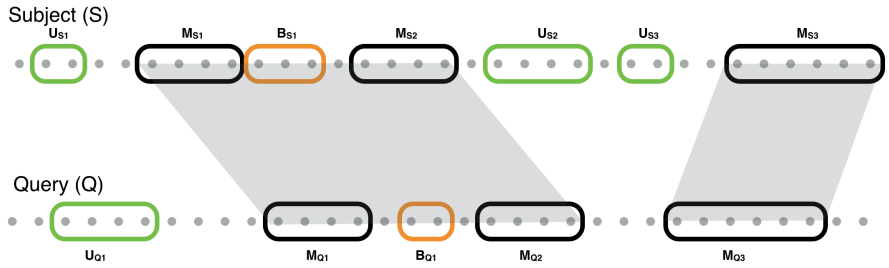


Figure B.18: Schematic of a match to illustrate the overall score calculation. Horizontal grey dotted lines denote amino acid sequences. The grey shaded areas are the matched regions. Light green rectangles represent unmatched features outside the matched regions, orange rectangles represent unmatched features inside the matched region. Black rectangles represent matching features inside the matched regions. Trig points are omitted for clarity.

B.4.3 Performance of the different scoring methods on test datasets

The scores that the light matter algorithm assigns should correspond with the structural reality. To investigate this behaviour I used the MinHashesScore, FeatureAAScore, and GreedySignificantBinScore scoring methods described in sections B.4.1 and B.4.2 to infer light matter scores for the matches in the 2HLA, 4MTP, and 4PH0 test datasets. I used the perfect finders[2], no trig point finders, and the HashFraction significance method with a significance cut-off of 0.01. Figures B.19a–c show the correlation between the Dali Z-scores and the scores computed by the different light matter scoring methods on the three test datasets.

Visually comparing the scatterplots for the scores computed by the FeatureAAScore (Fig. B.19a) and GreedySignificantBinScore scoring methods (Fig. B.19b) shows that they perform similarly, most likely due to the similarity in the score calculation (the median number of bins considered for the GreedySignificantBinScore calculation was 2). Due to the similar results of these two scoring methods, it seems best to only work with the FeatureAAScore, as its calculation involves less steps and is therefore faster to compute. Comparing the FeatureAAScore and MinHashesScore (Fig. B.19c), the scores computed by the FeatureAAScore scoring method correlate somewhat better with the Dali Z-scores. Therefore, I use the FeatureAAScore as the default scoring method in the light matter algorithm. However, neither scoring method reflects the structural reality very well, as highlighted by high light matter scores that are frequently assigned to matches with low (<8) Z-scores, and low light matter scores assigned to matches with high (>20) Z-scores.

B.4.4 Performance of the FeatureAAScore scoring method on two edge cases

Since the results from the previous section suggest that the scores computed by the FeatureAAScore scoring correlate best with the Dali Z-scores, I investigated its performance on two edge cases. On the one hand I investigated matches with a range of Z-scores that all have a light matter score above 0.99. On the other hand, I investigated the scores that are assigned to matches between structures that are not structurally similar, as measured by the Dali Z-score.

[2] 'PDB AlphaHelix_combined' and 'PDB ExtendedStrand'

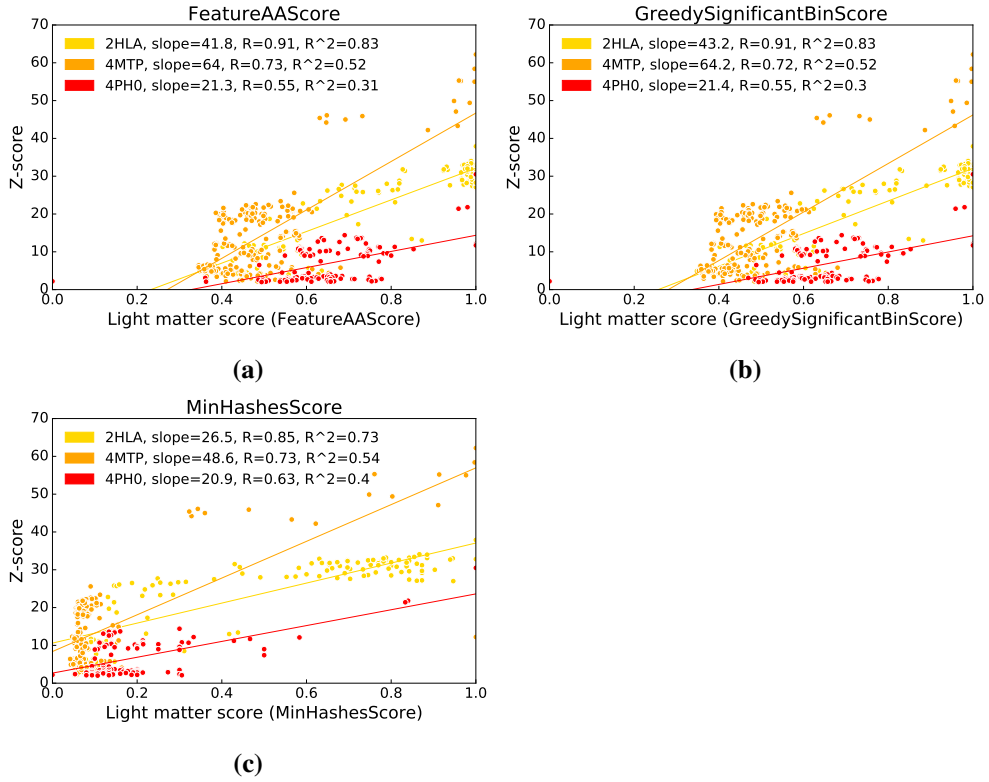


Figure B.19: Correlation of the MinHashesScore, FeatureAAScore and GreedySignificantBinScore methods with the Dali Z-score. Scores were calculated by applying the three scoring methods to the 2HLA, 4MTP, and 4PH0 test datasets, using the PDB AlphaHelix_combined and PDB ExtendedStrand finders, no trig points, and the HashFraction significance method with a significance cut-off of 0.01. Lines show the linear regression, with the shaded area showing the 95% confidence interval.

B.4.4.1 Investigating matches with light matter scores of 0.99 to 1.0

Figure B.20 shows the scores computed by the FeatureAAScore scoring method of the matches in the 4MTP dataset plotted against the Dali Z-scores. Interestingly, the plot shows a group of matches with Dali Z-scores between 62.2 to 47.1, which are assigned light matter scores of 0.99 or above (show in red). This would suggest that even though identical or very similar features are identified by the light matter algorithm, the tertiary structures are slightly different, leading to different Z-scores. Figure B.21 shows the structure comparisons and horizontal line plots of three examples of the matches described above. The example in Figs. B.21a–b) is the 4MTP:A sequence compared against itself, which has the highest Z-score of 62.2, and where we would expect the highest light matter score. Figures B.21c–d) show the 4MTP:A sequence against the 4MTP:D sequence, i.e., two different chains of the same pro-

tein compared against each other. The match is assigned a FeatureAAScore of 1.0 and a Z-score of 47.1. Looking at the horizontal line plots of those two matches, it is apparent that the type and order of features is the same, and the sequence is identical, while small variations in the structure lead to the lower Z-score. Figures B.21e–f) shows the influence of non-matching features on the FeatureAAScore calculation: looking at the horizontal line plot of the comparison between the 4MTP:A sequence and the 4K6M:B sequence, it is clear that the type and order of the features is the same in the two sequences, with the exception of a non-matching AC ExtendedStrand in 4MTP:A. This pulls the FeatureAAScore down to 0.99, a behaviour which is desirable and expected. Overall, in the cases of comparisons between very similar or identical sequences, the FeatureAAScore scoring method performs adequately.

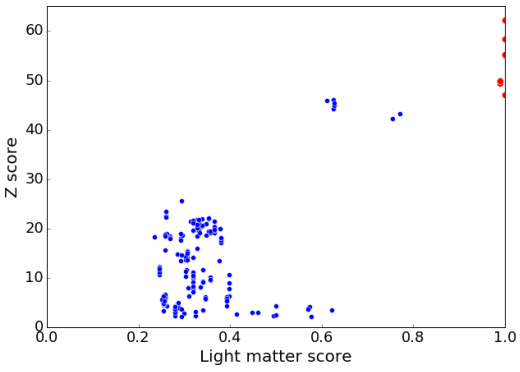


Figure B.20: Correlation of the FeatureAAScore and the Z-scores in the 4MTP dataset. I used the AC AlphaHelix_combined and AC ExtendedStrand landmark finders, no trig points, and the HashFraction significance method with a significance fraction of 0.01. Matches with a FeatureAAScore of 0.98 or better are plotted in red, other matches in blue.

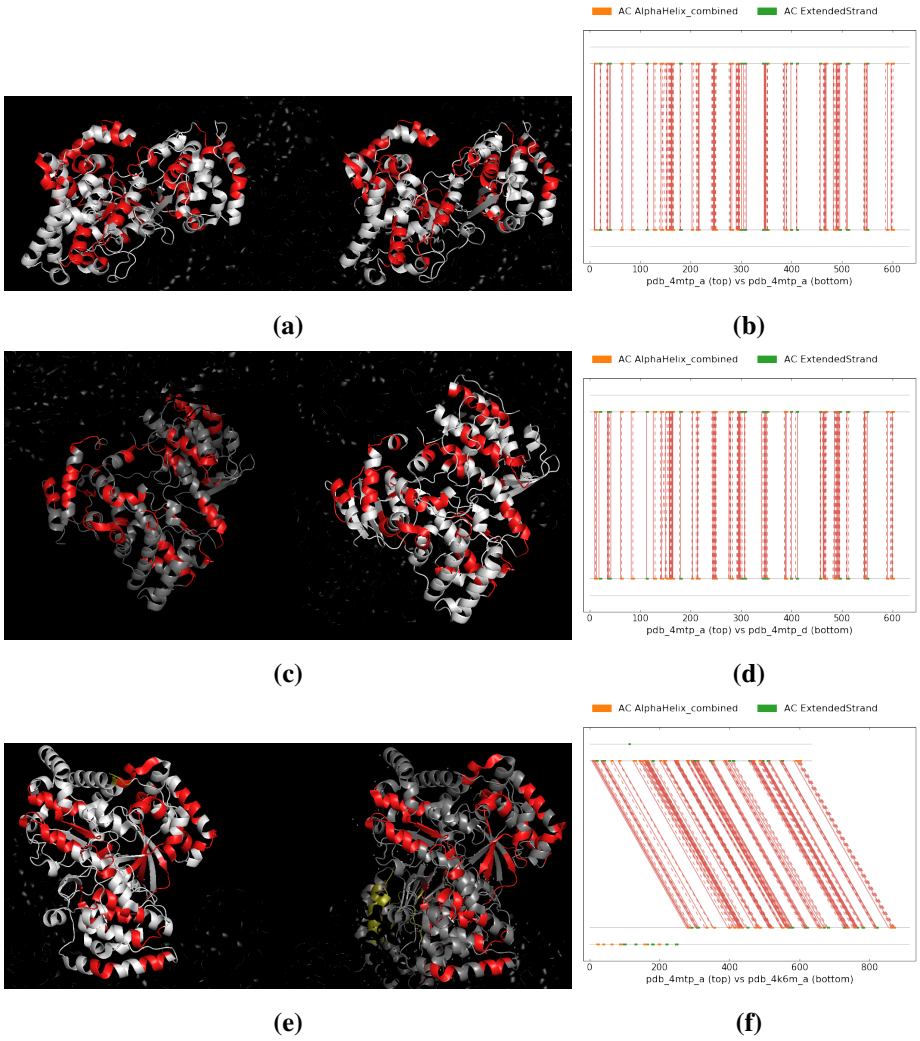


Figure B.21: Structure comparisons and horizontal line plots for matches with scores of 0.99 to 1.0. Each row of figures corresponds to one comparison. **a–b) Sequence 4MTP:A compared against itself, FeatureAAScore: 1.0, Z-score: 62.2. c–d) Sequence 4MTP:A compared against sequence 4MTP:D, FeatureAAScore: 1.0, Z-score: 47.1. e–f) Sequence 4MTP:A (left) compared against sequence 4K6M:B (right), FeatureAAScore: 0.99, Z-score: 49.4.** Figures a, c, and e) show the comparison of the two structures, with 4MTP:A always on the right. Matching features are shown in red, non-matching features in muddy green. Figures b, d, and f) show the horizontal line plots for the comparisons in questions, with 4MTP:A always being the top sequence. Only the best bin is shown.

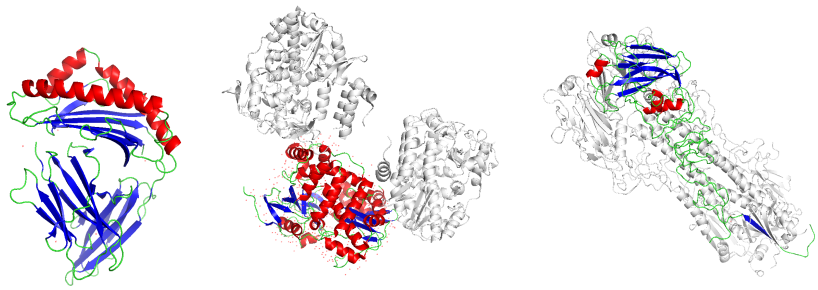
B.4.4.2 Investigating scores of matches between dis-similar structures

The second edge case I investigated was how the FeatureAAScore scoring method performs on sequences that correspond to structures that are not homologous. This is to get a basic understanding of which light matter scores we may have to expect for false positive matches. To this end, I compared the 2HLA:A sequence against all sequences in the Polymerase and HA test datasets, using the 'PDB AlphaHelix_combined' and 'PDB ExtendedStrand' landmark finders. For a visual impression of the 2HLA:A structure, as well as an example of a structure from the Polymerase and HA test dataset, see Fig. B.22. Figures B.23a) and b) show the scores computed by the FeatureAAScore scoring method resulting from the comparison of 2HLA:A sequence against all sequences in the Polymerase and HA test datasets plotted against the BLAST bit scores (Dali is not able to detect any structural similarity, so only the BLAST bit scores are shown). The bit scores are in a range that can be expected from non-matching sequences. The light matter scores on the other hand, are surprisingly high (between 0.3 and 0.6), given that I do not expect any sequences to match between those test datasets. Looking at Fig. B.22, one could argue that there is some similarity in the arrangement of the secondary structures between the 2HLA:A and the HA domain that forms part of the match, but the same argument cannot be made for the Polymerase, and the light matter scores are similar for both Polymerase and HA test datasets. Two comparisons with the highest FeatureAAScores, one from each test dataset, were investigated further. Each row in Fig. B.24 shows the three-dimensional structures (left) and horizontal line plot (right) of a comparison. The top row shows the comparison of the 2HLA:A sequence against the 1N35:A sequence (from the Polymerase test dataset) and the bottom row showing the comparison of the 2HLA:A sequence against the 5HMG:E sequence (from the HA test dataset). When looking at Figs. B.24a) and c), it becomes apparent that while there are matching pairs between the two structures (shown in red, connected by white lines), they do not correspond structurally and it does not indicate that there is structural similarity. The horizontal line plot in Fig. B.24b indicates that the high score between the 2HLA:A and 1N35:A sequence is most likely caused by the length normaliser employed in the FeatureAAScore calculation, which rewards the situation where the matched region that spans almost the entire length of at least one of the two sequences. Indeed, the factor for the length normaliser for the match in Fig. B.24b is 0.86. The high score assigned to the match between 2HLA:A and 5HMG:E sequences seems to be caused by the algorithm identifying spurious matches between a number of secondary structures present in both sequences. This suggest that the comparison of sequences based

on their secondary structures may not be able to accurately differentiate between structures that are different at the tertiary structure level.

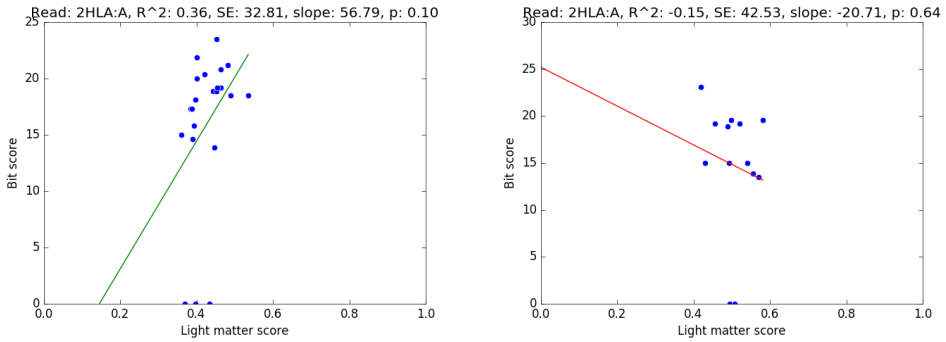
B.4.5 Conclusion

Overall, the scores computed by the three scoring methods of the light matter algorithm correlate with the Dali Z-scores (Fig. B.19). However, neither scoring method reflects the structural reality for all pairwise comparisons in the test datasets, as sometimes high light matter scores are computed for matches with low (<8) Z-scores, and low light matter scores for matches with high (>20) Z-scores. While the FeatureAAScore scoring method performs adequately for highly similar sequences, it assigns relatively high scores (0.3–0.6) to matches between sequences coding for proteins with dis-similar structures. This is partly due to the length normalisation performed in the score calculation, but also suggests that in some cases, the similarity between secondary structures of two unrelated sequences may lead to false positive matches. Therefore, care needs to be taken when determining which sequences are actually considered to match (and are therefore structurally equivalent), and which scores are indicative of a good match, both of which are areas that require further work.



(a) **(b)** **(c)**

Figure B.22: Comparison of the 2HLA:A structure (a)), the 3UQS:A structure from the Polymerase dataset (b)), and the 5HMG:E structure from the HA dataset (c)). Helices are shown in red, strands in blue, chains that represent the respective sequence in light green, and chains not used in the light matter comparison in white. The size of the three structures is not scaled the same.



(a) **(b)**

Figure B.23: Correlation of the bit score and the FeatureAAScore light matter score. The 'PDB AlphaHelix_combined' and 'PDB ExtendedStrand' landmark finders were used, no trig points, and the HashFraction significance method with a significance fraction of 0.01. **a) 2HLA:A against Polymerase test dataset. b) 2HLA:A against HA test dataset.** The lines show the linear regression in green (positive correlation) and red (negative correlation).

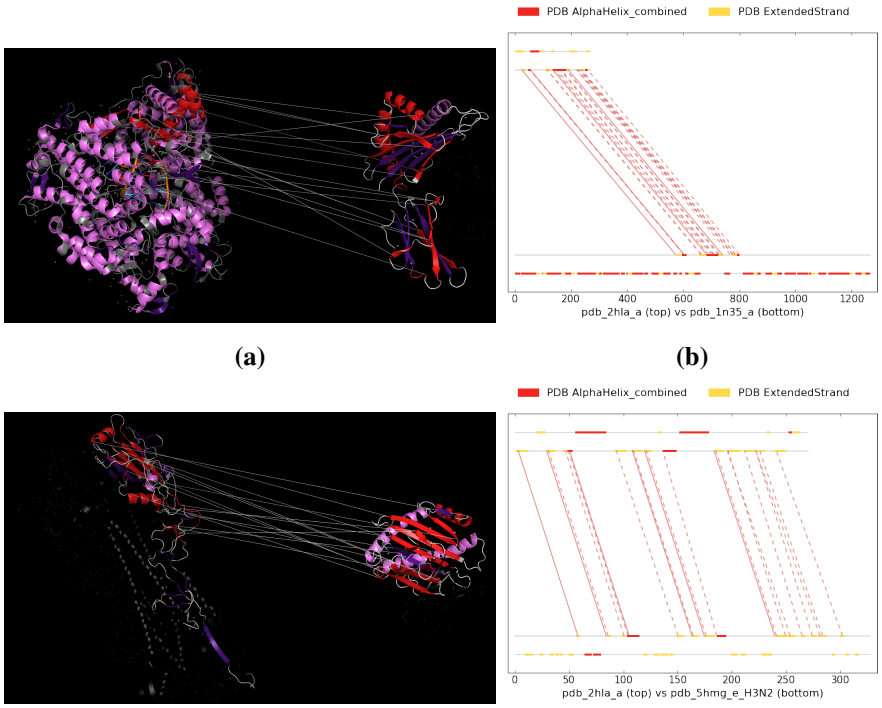


Figure B.24: Structure comparisons and horizontal line plots for matches of unrelated sequences. Each row of figures corresponds to one comparison. **a–b) 2HLA:A against 1N35:A (Polymerase test dataset), FeatureAAScore: 0.54, Bit score: 18.5, c–d) 2HLA:A against 5HMG:E (HA test dataset), FeatureAAScore: 0.58, Bit score: 19.6.** Figures a) and c) show the comparison of the two structures, with 2HLA:A on the right. Matching features are shown in red, non-matching helices in light purple, non-matching extended strands in dark purple. Matching features in the two structures are connected by white lines. Figures b) and d) show the horizontal line plots for the comparisons in questions, with 2HLA:A being the top sequence. Only the best bin is shown. No structural similarity as measured by the Dali Z-score could be detected.

APPENDIX C: LIGHT MATTER ALGORITHM: PRELIMINARY RESULTS

During the development of the light matter algorithm, I periodically tested it on some or all of the five test datasets (chapter B.1.2), to better understand the effects of the functionalities we implemented. This chapter is an extension of the results presented in chapter 8 of this thesis. In the first section, I show a general overview of the performance of the algorithm under different parameter combinations, using correlations between light matter scores and Dali Z-scores. The second section aims to improve our understanding some of the patterns observed in the light matter scoring based on visualisations of the distribution of features on the three-dimensional protein structures.

C.1 EVALUATION OF DIFFERENT FINDER AND PARAMETER COMBINATIONS

Chapter B.2 described the development of the structural features and parameters that are used by the light matter algorithm. In this chapter, I present an evaluation of the effects when those finders and parameters are used concurrently. I tested the following parameter combinations on the five test datasets:

- Performance the light matter algorithm using different combinations of landmark finders, and the 'Peak', 'Trough', and 'AminoAcids' trig point finders (Fig. C.1).
- Performance the light matter algorithm using different combinations of landmark finders, and no trig point finders (Fig. C.2).
- Performance the light matter algorithm using different combinations of trig point finders, and the 'PDB AlphaHelix', 'PDB AlphaHelix_3_10', 'PDB AlphaHelix_pi', and 'PDB ExtendedStrand' landmark finders (Fig. C.3).
- Performance the light matter algorithm using different values for the FeatureLengthBase parameter (Fig. C.4).
- Performance the light matter algorithm using different values for the DistanceBase parameter (Fig. C.5).

– Performance the light matter algorithm using different values for the MaxDistance parameter (Fig. C.6).

I used the FeatureAAScore scoring method and HashFraction significance method with a significance fraction of 0.01 throughout. In Figs. C.1–C.6, rows correspond to different test datasets (top to bottom: 2HLA, 4MTP, HA, 4PH0, Polymerase), and columns correspond to a different set of parameters under investigation. Each plot shows the light matter scores computed with the relevant parameters plotted against the Dali Z-scores of that particular match. Lines indicate the linear regression fitted to the scores, with its parameters indicated in the title of the subfigure.

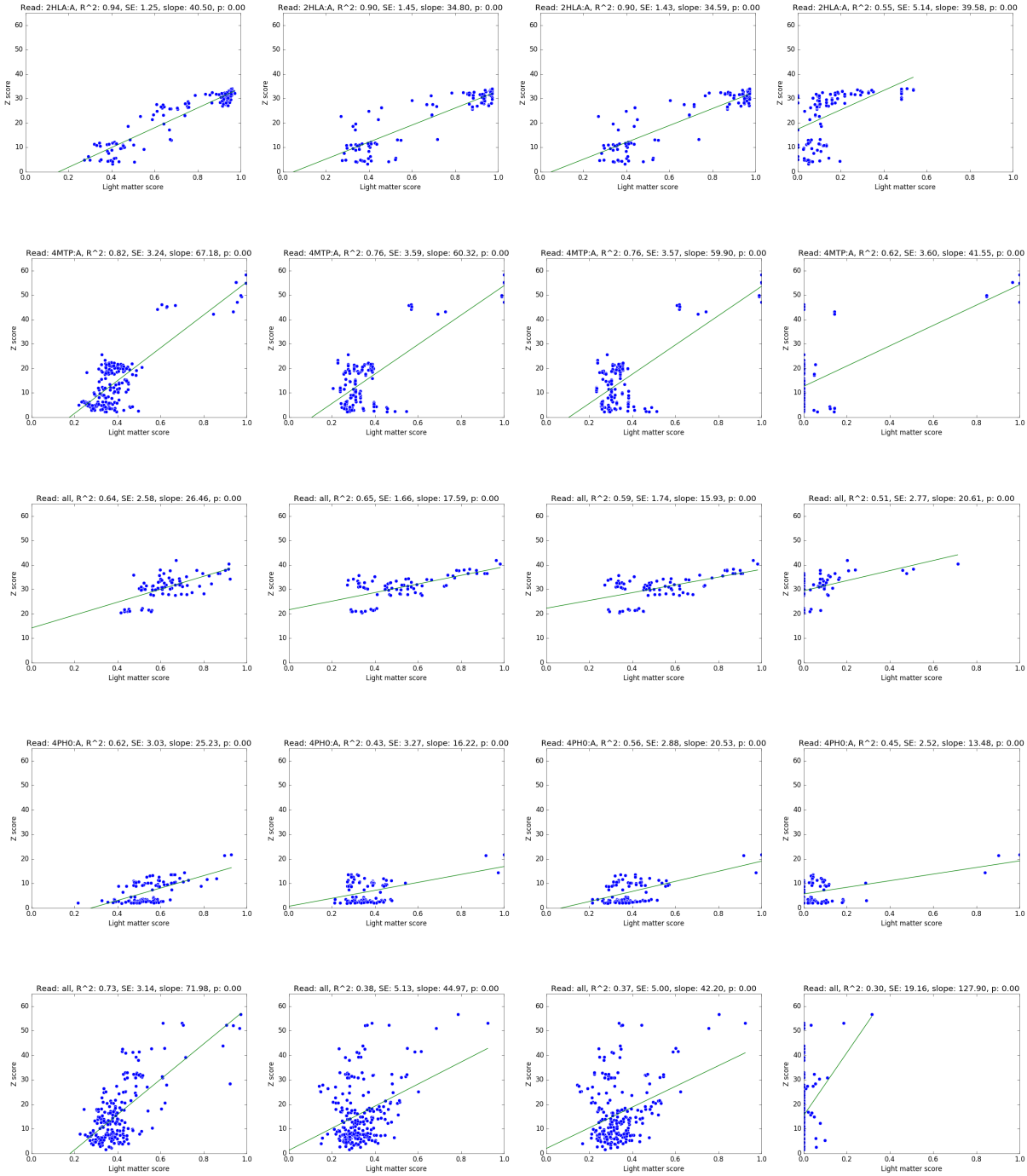
Figure C.1 shows the light matter scores inferred with different landmark finders and the same set of trig point finders ('Peak', 'Trough', and 'AminoAcids'), plotted against the Dali Z-scores. The first column shows the scores calculated using the 'PDB AlphaHelix', 'PDB AlphaHelix_3_10', 'PDB AlphaHelix_pi', and 'PDB ExtendedStrand' finders, for reference. Based on the linear regression fitted to the light matter and Dali Z-scores, neither of the two different combinations of Aho-Corasick finders in columns two ('AC AlphaHelix', 'AC AlphaHelix_3_10', 'AC AlphaHelix_pi', 'AC ExtendedStrand') and three ('AC AlphaHelix_combined', 'AC ExtendedStrand') performs consistently better. Furthermore, landmark finders based on the GOR4 algorithm shown in column four ('GOR4 AlphaHelix', 'GOR4 BetaStrand', and 'GOR4 Coil') lead to a large number of matches with light matter scores of 0.0. Therefore, using these finders without any additional landmark finders is not desirable. The same is true for the combination of 'EukaryoticLinearMotif', 'AminoAcidsLm' and 'Prosite' landmark finders (column seven). This may be explained by the number of features that each set of landmark finders identifies in the five test datasets: the GOR4 finders combined identify 27,757 features, 'EukaryoticLinearMotif', 'AminoAcidsLm', and 'Prosite' finders combined identify 37,668 features, while the Aho-Corasick finders identify 56,550 (using 'AC AlphaHelix_combined') or 55,404 (using 'AC AlphaHelix', 'AC AlphaHelix_3_10', and 'AC AlphaHelix_pi separately) features. More features may make it more likely for a match to be identified as significant, and hence be assigned a score > 0.0. The combination of the 'AC AlphaHelix_combined', 'AC ExtendedStrand', 'Prosite', 'EukaryoticLinearMotif', and 'AminoAcidsLm' landmark finders in column six performs best, taking into account the correlation coefficients of the linear regression, as well as the low number of matches that are assigned a light matter score of 0.0.

Figure C.2 is structured identically as Fig. C.1. It shows the evaluations for the same combinations of landmark finders as Fig. C.1, but without trig point finders. Comparing Figs. C.1 and C.2, the evaluations that include trig points in Fig. C.1

consistently have a better correlation of the light matter score and Dali Z-scores, as indicated by the linear regression, than when no trig points were used, suggesting that the 'Peak', 'Trough', and 'AminoAcids' trig points improve the performance of the light matter algorithm.

Figure C.3 shows the performance of the different trig point finders, while using the same combination of landmark finders ('PDB AlphaHelix', 'PDB AlphaHelix_3_10', 'PDB AlphaHelix_pi', and 'PDB ExtendedStrand'). There is very little difference in the light matter scores computed for different combinations of trig point finders. When considering the percentage of pairs that are formed between a landmark and a trig point, as opposed to a landmark and a landmark in the five test dataset, only 16.6% of all pairs consist of a landmark and a trig point. This is because a landmark is always first paired with other landmarks, and then with any trig points. Based on the correlation coefficients of the linear regression, using either both 'Peaks' and 'Troughs', or 'Peaks', 'Troughs', and 'AminoAcids' trig point finders together leads to the best correlation of light matter scores and Dali Z-scores.

C. LIGHT MATTER ALGORITHM: PRELIMINARY RESULTS



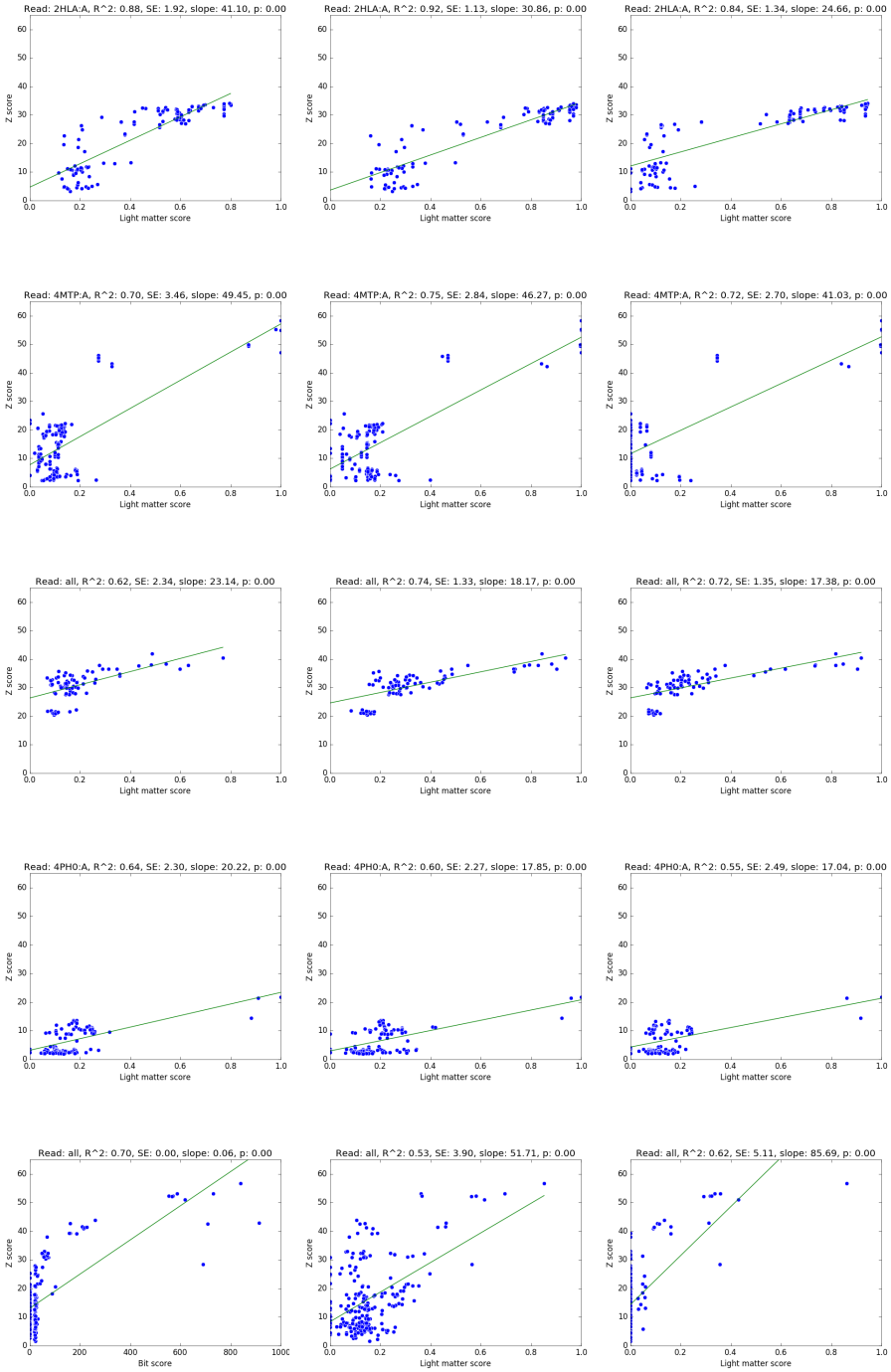
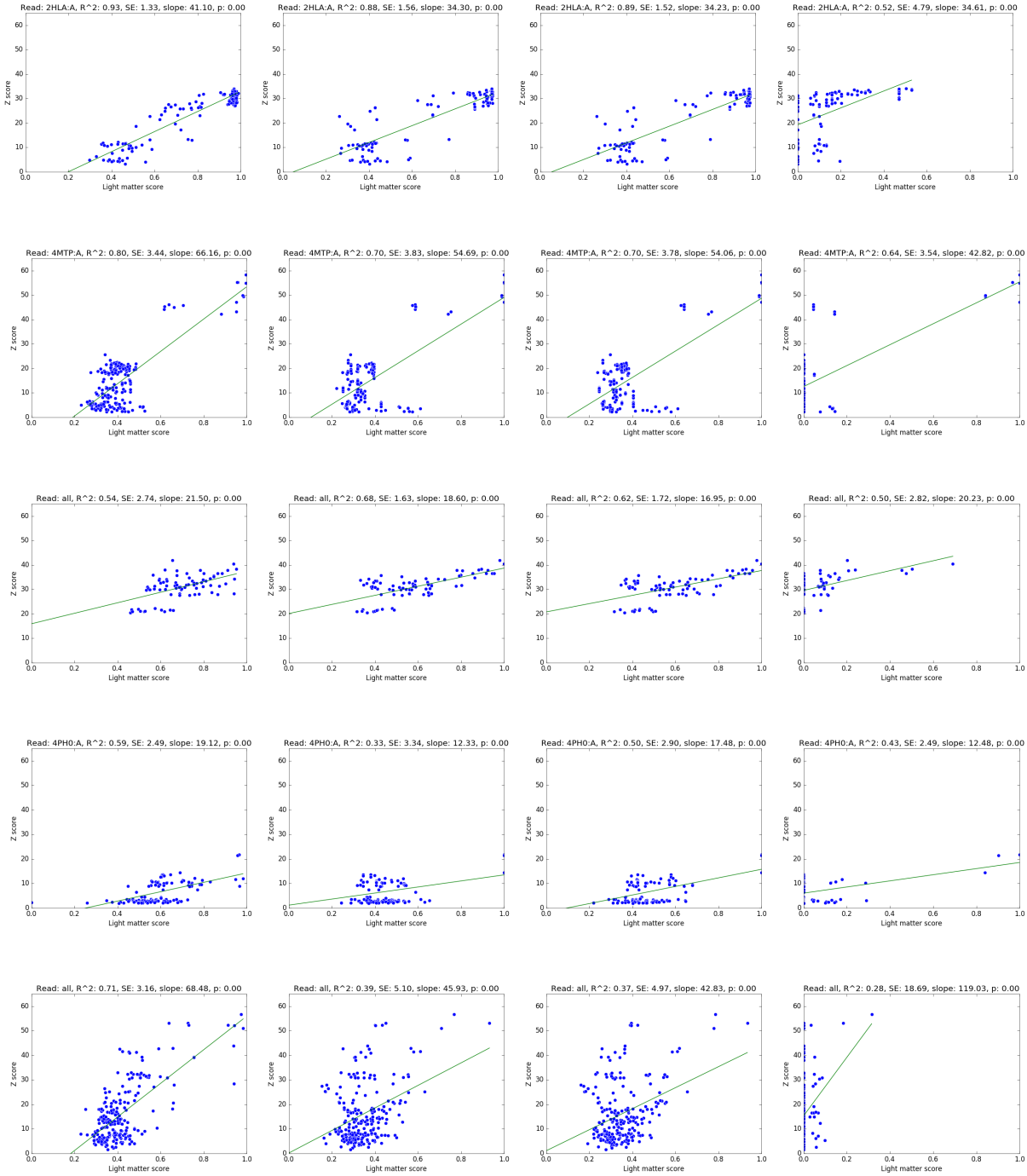

Read: 2HLA:A, R^2: 0.88, SE: 1.92, slope: 41.10, p: 0.00
Read: 2HLA:A, R^2: 0.92, SE: 1.13, slope: 30.86, p: 0.00
Read: 2HLA:A, R^2: 0.84, SE: 1.34, slope: 24.66, p: 0.00
Read: 4MTP:A, R^2: 0.70, SE: 3.46, slope: 49.45, p: 0.00
Read: 4MTP:A, R^2: 0.75, SE: 2.84, slope: 46.27, p: 0.00
Read: 4MTP:A, R^2: 0.72, SE: 2.70, slope: 41.03, p: 0.00
Read: all, R^2: 0.62, SE: 2.34, slope: 23.14, p: 0.00
Read: all, R^2: 0.74, SE: 1.33, slope: 18.17, p: 0.00
Read: all, R^2: 0.72, SE: 1.35, slope: 17.38, p: 0.00
Read: 4PH0:A, R^2: 0.64, SE: 2.30, slope: 20.22, p: 0.00
Read: 4PH0:A, R^2: 0.60, SE: 2.27, slope: 17.85, p: 0.00
Read: 4PH0:A, R^2: 0.55, SE: 2.49, slope: 17.04, p: 0.00
Read: all, R^2: 0.70, SE: 0.00, slope: 0.06, p: 0.00
Read: all, R^2: 0.53, SE: 3.90, slope: 51.71, p: 0.00
Read: all, R^2: 0.62, SE: 5.11, slope: 85.69, p: 0.00
Z score
Light matter score
Bit score

Figure C.1: Correlation of FeatureAAScore and Z-scores, using different combinations of landmark finders. Each row corresponds to a different test dataset, top to bottom: 1) 2HLA. 2) 4MTP. 3) HA. 4) 4PH0. 5) Polymerase. Each column corresponds to a different combination of landmark finders, left to right: 1) 'PDB AlphaHelix', 'PDB AlphaHelix_3_10', 'PDB AlphaHelix_pi', and 'PDB ExtendedStrand'. 2) 'AC AlphaHelix', 'AC AlphaHelix_3_10', 'AC AlphaHelix_pi', and 'AC ExtendedStrand'. 3) 'AC AlphaHelix_combined', and 'AC ExtendedStrand'. 4) 'GOR4 AlphaHelix', 'GOR4 BetaStrand', and 'GOR4 Coil'. 5) 'AC AlphaHelix_combined', 'AC ExtendedStrand', 'GOR4 AlphaHelix', 'GOR4 BetaStrand', and 'GOR4 Coil'. 6) 'AC AlphaHelix_combined', 'AC ExtendedStrand', 'Prosite', 'EukaryoticLinearMotif', and 'AminoAcidsLm'. 7) 'Prosite', 'EukaryoticLinearMotif', and 'AminoAcidsLm'. The 'Peak', 'Trough', and 'AminoAcids' trig point finders, a significance fraction of 0.01, a FeatureLengthBase of 1.01, DistanceBase of 1.1, DeltaScale of 1.0, and a MaxDistance of 1000 were used throughout. Lines indicate the linear regression, with coefficients in the title of each subfigure.

C. LIGHT MATTER ALGORITHM: PRELIMINARY RESULTS



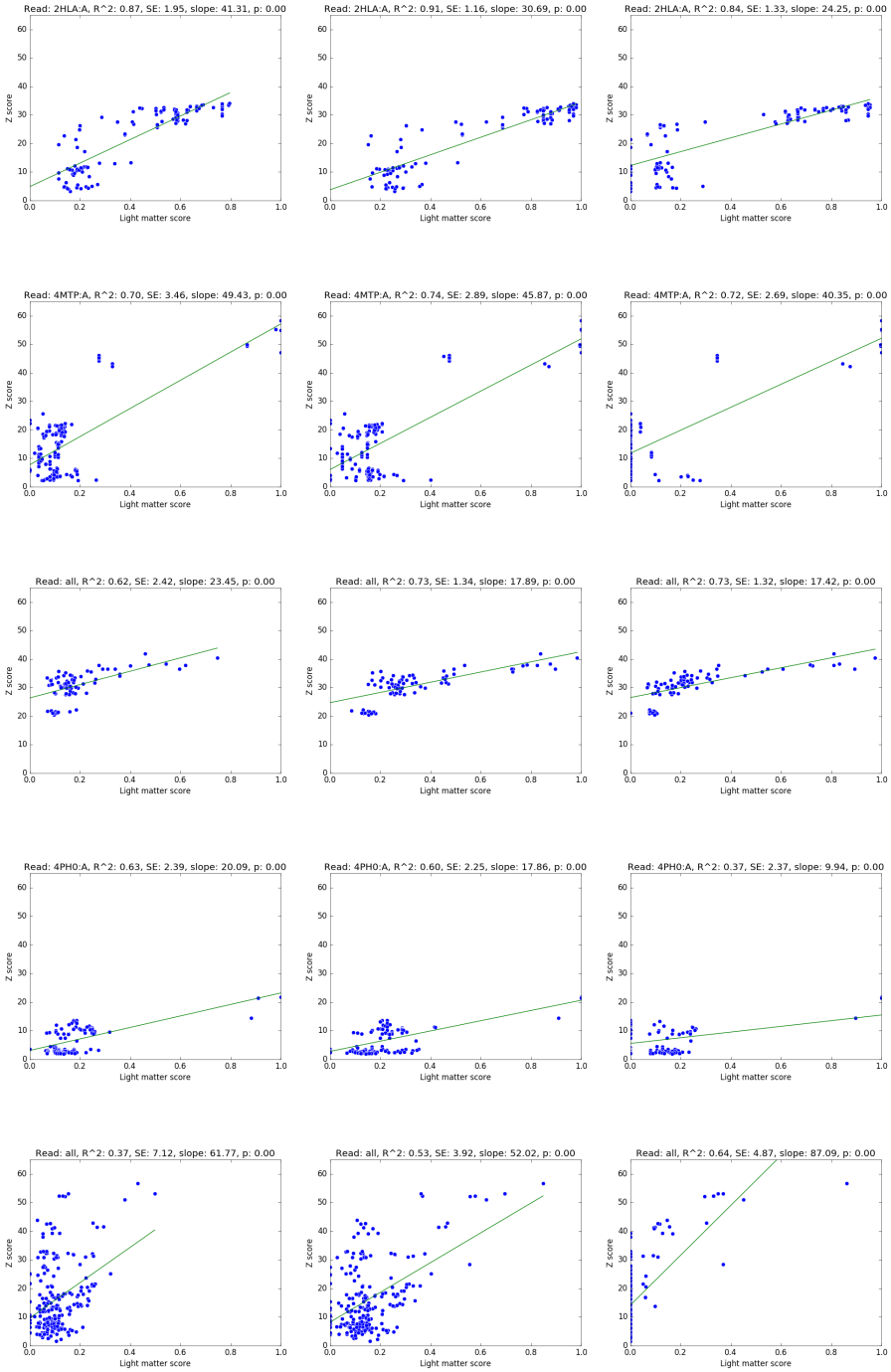
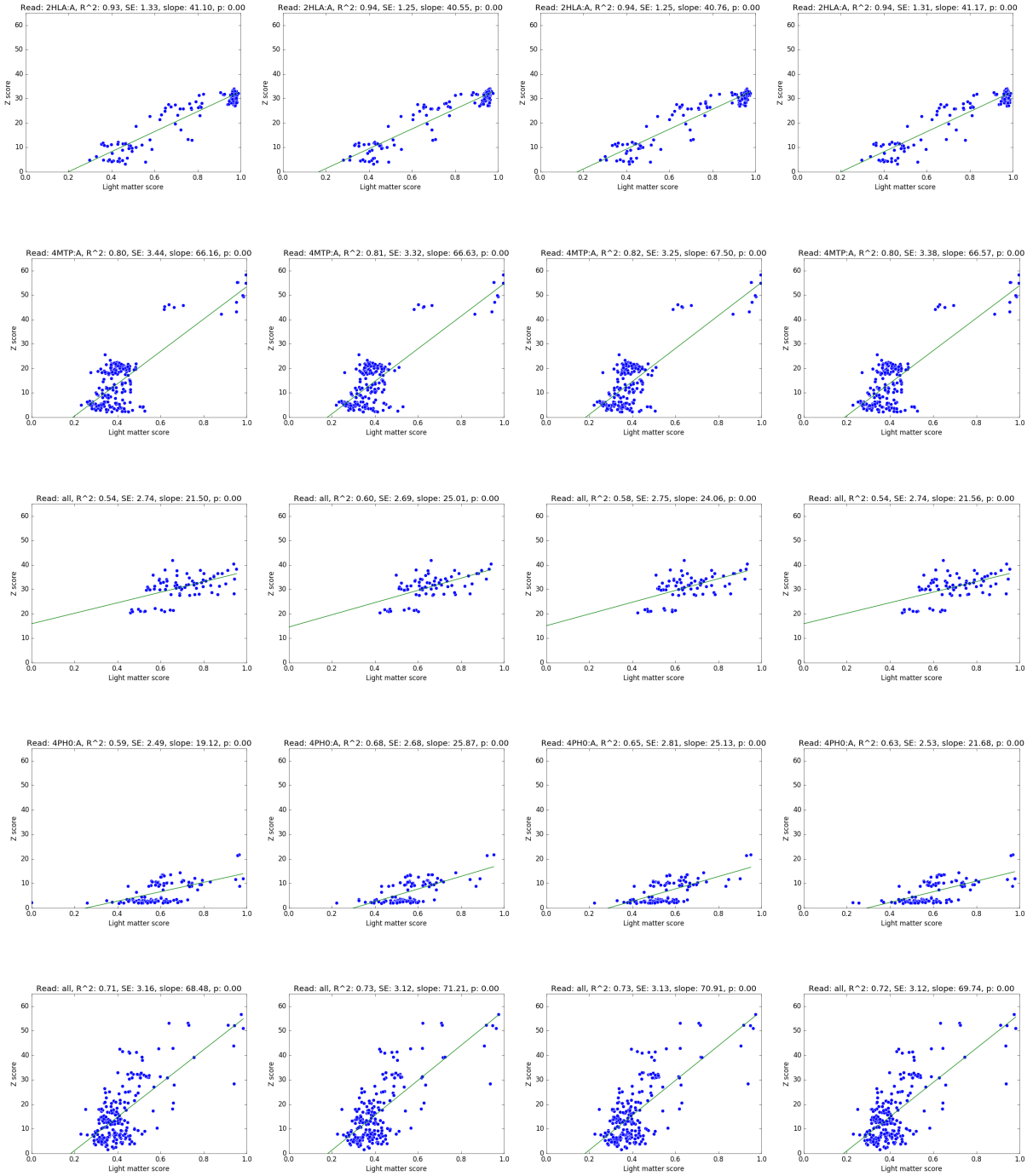

Read: 2HLA:A, R^2: 0.87, SE: 1.95, slope: 41.31, p: 0.00
Read: 2HLA:A, R^2: 0.91, SE: 1.16, slope: 30.69, p: 0.00
Read: 2HLA:A, R^2: 0.84, SE: 1.33, slope: 24.25, p: 0.00
Read: 4MTP:A, R^2: 0.70, SE: 3.46, slope: 49.43, p: 0.00
Read: 4MTP:A, R^2: 0.74, SE: 2.89, slope: 45.87, p: 0.00
Read: 4MTP:A, R^2: 0.72, SE: 2.69, slope: 40.35, p: 0.00
Read: all, R^2: 0.62, SE: 2.42, slope: 23.45, p: 0.00
Read: all, R^2: 0.73, SE: 1.34, slope: 17.89, p: 0.00
Read: all, R^2: 0.73, SE: 1.32, slope: 17.42, p: 0.00
Read: 4PH0:A, R^2: 0.63, SE: 2.39, slope: 20.09, p: 0.00
Read: 4PH0:A, R^2: 0.60, SE: 2.25, slope: 17.86, p: 0.00
Read: 4PH0:A, R^2: 0.37, SE: 2.37, slope: 9.94, p: 0.00
Read: all, R^2: 0.37, SE: 7.12, slope: 61.77, p: 0.00
Read: all, R^2: 0.53, SE: 3.92, slope: 52.02, p: 0.00
Read: all, R^2: 0.64, SE: 4.87, slope: 87.09, p: 0.00
Z score
Light matter score

Figure C.2: Correlation of FeatureAAScore and Z-scores, using different combinations of landmark- and no trig point finders. Rows, top to bottom: 1) 2HLA. 2) 4MTP. 3) HA. 4) 4PH0. 5) Polymerase. Columns, left to right: 1) 'PDB AlphaHelix', 'PDB AlphaHelix_3_10', 'PDB AlphaHelix_pi', and 'PDB ExtendedStrand'. 2) 'AC AlphaHelix', 'AC AlphaHelix_3_10', 'AC AlphaHelix_pi', and 'AC ExtendedStrand'. 3) 'AC AlphaHelix_combined', and 'AC ExtendedStrand'. 4) 'GOR4 AlphaHelix', 'GOR4 BetaStrand', and 'GOR4 Coil'. 5) 'AC AlphaHelix_combined', 'AC ExtendedStrand', 'GOR4 AlphaHelix', 'GOR4 BetaStrand', and 'GOR4 Coil'. 6) 'AC AlphaHelix_combined', 'AC ExtendedStrand', 'Prosite', 'EukaryoticLinearMotif', and 'AminoAcidsLm'. 7) 'Prosite', 'EukaryoticLinearMotif', and 'AminoAcidsLm'. A significance fraction of 0.01, no trig points, a FeatureLengthBase of 1.01, DistanceBase of 1.1, DeltaScale of 1.0, and a MaxDistance of 1000 were used throughout. Lines indicate the linear regression, with coefficients in the title of each subfigure.

C. LIGHT MATTER ALGORITHM: PRELIMINARY RESULTS



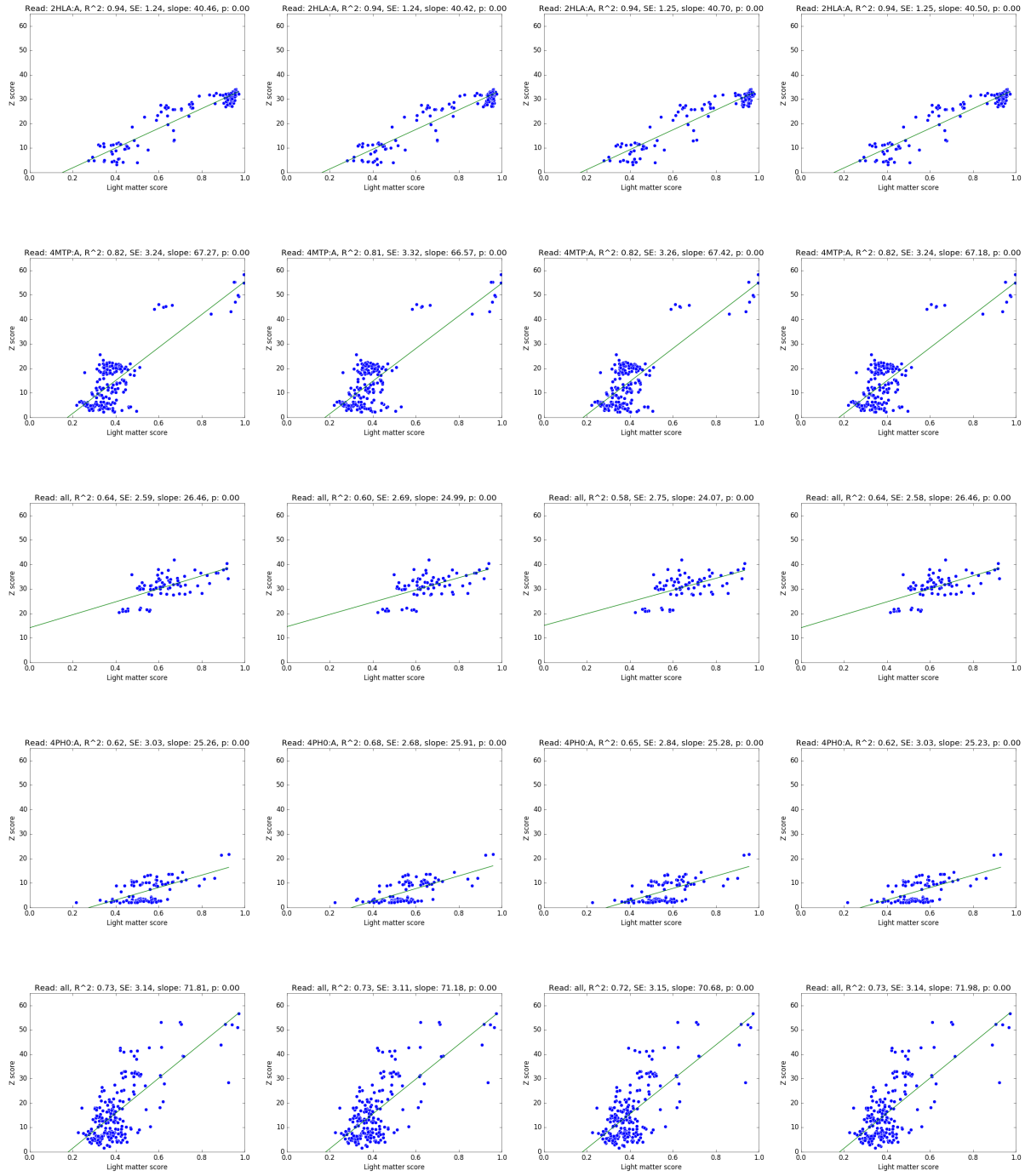

Read: 2HLA:A, R^2: 0.94, SE: 1.24, slope: 40.46, p: 0.00
Read: 2HLA:A, R^2: 0.94, SE: 1.24, slope: 40.42, p: 0.00
Read: 2HLA:A, R^2: 0.94, SE: 1.25, slope: 40.70, p: 0.00
Read: 2HLA:A, R^2: 0.94, SE: 1.25, slope: 40.50, p: 0.00
Read: 4MTP:A, R^2: 0.82, SE: 3.24, slope: 67.27, p: 0.00
Read: 4MTP:A, R^2: 0.81, SE: 3.32, slope: 66.57, p: 0.00
Read: 4MTP:A, R^2: 0.82, SE: 3.26, slope: 67.42, p: 0.00
Read: 4MTP:A, R^2: 0.82, SE: 3.24, slope: 67.18, p: 0.00
Read: all, R^2: 0.64, SE: 2.59, slope: 26.46, p: 0.00
Read: all, R^2: 0.60, SE: 2.69, slope: 24.99, p: 0.00
Read: all, R^2: 0.58, SE: 2.75, slope: 24.07, p: 0.00
Read: all, R^2: 0.64, SE: 2.58, slope: 26.46, p: 0.00
Read: 4PH0:A, R^2: 0.62, SE: 3.03, slope: 25.26, p: 0.00
Read: 4PH0:A, R^2: 0.68, SE: 2.68, slope: 25.91, p: 0.00
Read: 4PH0:A, R^2: 0.65, SE: 2.84, slope: 25.28, p: 0.00
Read: 4PH0:A, R^2: 0.62, SE: 3.03, slope: 25.23, p: 0.00
Read: all, R^2: 0.73, SE: 3.14, slope: 71.81, p: 0.00
Read: all, R^2: 0.73, SE: 3.11, slope: 71.18, p: 0.00
Read: all, R^2: 0.72, SE: 3.15, slope: 70.68, p: 0.00
Read: all, R^2: 0.73, SE: 3.14, slope: 71.98, p: 0.00
Z score
Light matter score

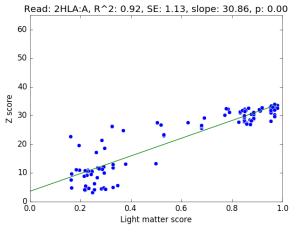
Figure C.3: Correlation of FeatureAAScore and the Z-scores, using different combinations of trig point finders. Rows, top to bottom: 1) 2HLA. 2) 4MTP. 3) HA. 4) 4PH0. 5) Polymerase. Columns, left to right: 1) no trig points. 2) 'Peaks'. 3) 'Troughs'. 4) 'AminoAcids'. 5) 'Peaks', 'Troughs'. 6) 'Peaks', 'AminoAcids'. 7) 'Troughs', 'AminoAcids'. 8) 'Peaks', 'Troughs', 'AminoAcids'. A significance fraction of 0.01, the 'PDB AlphaHelix', 'PDB AlphaHelix_3_10', 'PDB AlphaHelix_pi', and 'PDB ExtendedStrand' landmark finders, a FeatureLengthBase of 1.01, DistanceBase of 1.1, DeltaScale of 1.0, and a MaxDistance of 1000 were used throughout. Lines indicate the linear regression, with coefficients in the title of each subfigure.

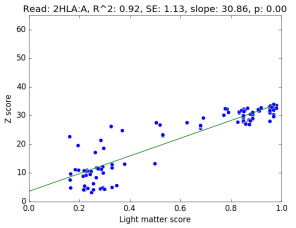
In addition to different combinations of landmark and trig point finders, I also tested different values for the FeatureLenghtBase, DistanceBase, and MaxDistance parameters in the light matter algorithm. The FeatureLenghtBase is the logarithmic scaling factor applied to the length of the landmarks that are longer than one, the DistanceBase logarithmically scales the distances between the two features in a pair, and the MaxDistance is the maximum distance that two features in a pair can be apart in order to be paired, measured in amino acids.

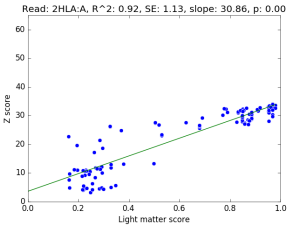
Figure C.4 shows the effect of different values for the FeatureLenghtBase parameter. Comparing the regressions, feature length scaling only takes effect when a FeatureLenghtBase of 1.35 or above is used. This can be explained by the lengths of the landmark features that I used: the substrings used by the 'AC AlphaHelix_combined' finder range in length from 4 to 13 amino acids, with a median of 6, while the 'AC ExtendedStrand' features range from 4 to 9 amino acids, with a median of 5 (Fig. B.8b). Due to the logarithmic scaling, the scaling of feature lengths only takes effect when features longer than four amino acids are scaled with a featureLengthBase of 1.35 (Fig. 7.2). Thus, on short features, scaling will not take effect. The median length of 'AC AlphaHelix_combined' and 'AC ExtendedStrand' features in the five test datasets is 5, which could be the reason for the limited effect of the FeatureLengtBase parameter. Based on these evaluations, the effect of scaling feature lengths is negligible.

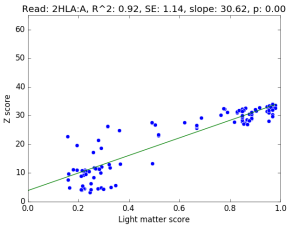
Figure C.5 shows the performance of different values for the DistanceBase parameter. Based on the results of the linear regression, in our five test datasets, scaling of the distances between features in pairs only starts having an effect when a parameter value of 1.1 or above is used. Visually comparing the scatter plots of the light matter and Z-scores shows that higher values for the DistanceBase parameter lead to higher light matter scores, without shifting their values relative to each other, up to a DistanceBase of 1.35. The increase in the value of the DistanceBase parameter allows more pairs to match, leading to higher values of the numerator in the MatchedRegionScore in the FeatureAAScore calculation, and hence to the relative increase of the light matter scores. Based on the correlation coefficients, using a DistanceBase of 1.35 results in the best correlation between light matter scores and Dali Z-scores.

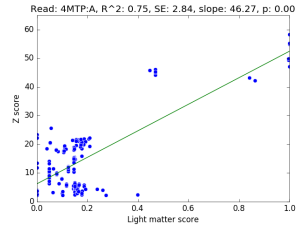
C. LIGHT MATTER ALGORITHM: PRELIMINARY RESULTS

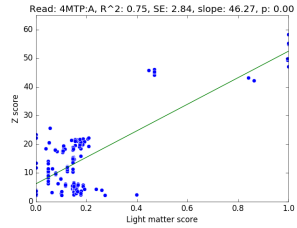


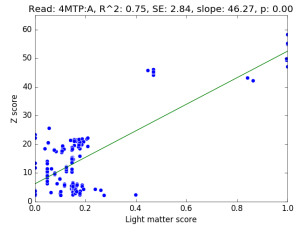


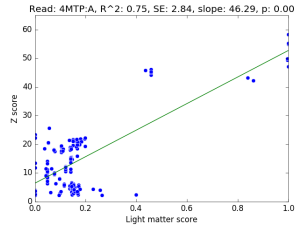


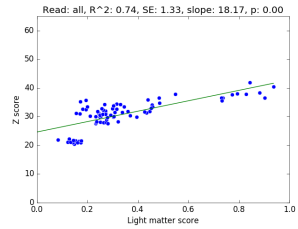


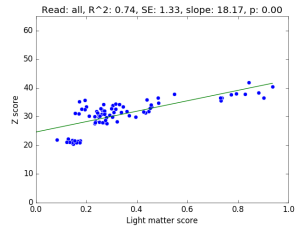


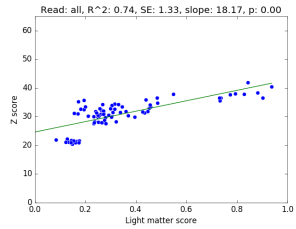


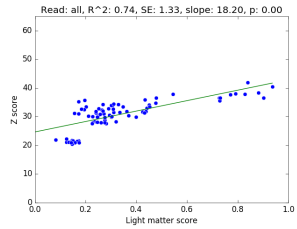


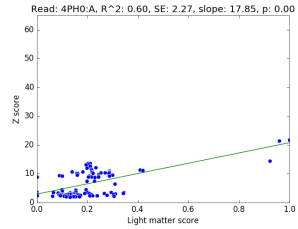


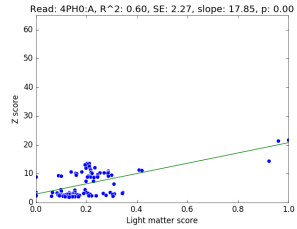


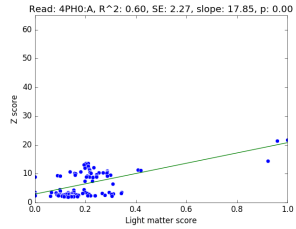


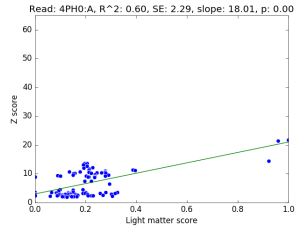


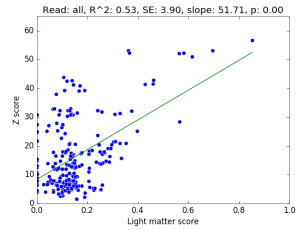


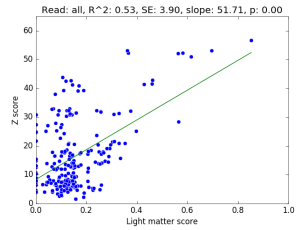


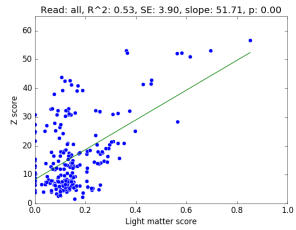












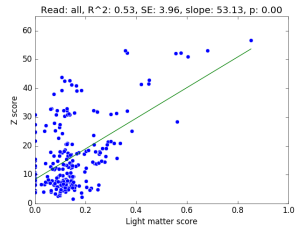
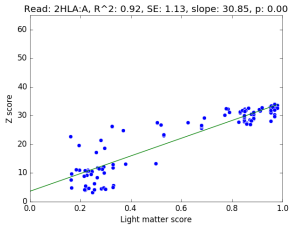
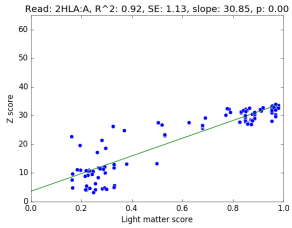
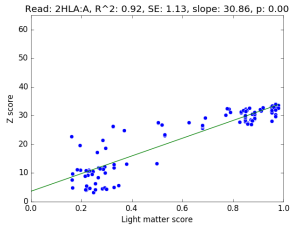


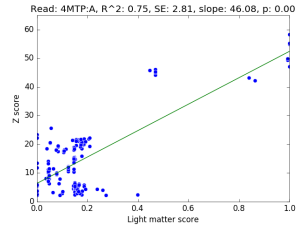
Figure C.4: Correlation of FeatureAAScore and Z-scores, using different values for the FeatureLengthBase parameter. Rows, top to bottom: 1) 2HLA. 2) 4MTP. 3) HA. 4) 4PH0. 5) Polymerase. Columns, left to right: 1) FeatureLengthBase: 1.0. 2) FeatureLengthBase: 1.1. 3) FeatureLengthBase: 1.2. 4) FeatureLengthBase: 1.35. 'AC AlphaHelix_combined', 'AC ExtendedStrand', 'Prosite', 'EukaryoticLinearMotif', and 'AminoAcidsLm' landmark finders and 'Peak', 'Trough', and 'AminoAcids' trig point finders, a significance fraction of 0.01, a DistanceBase of 1.1, DeltaScale of 1.0, and a MaxDistance of 1000 were used throughout. Lines indicate the linear regression, with coefficients in the title of each subfigure.

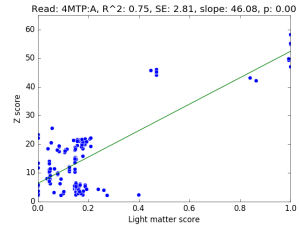
C. LIGHT MATTER ALGORITHM: PRELIMINARY RESULTS

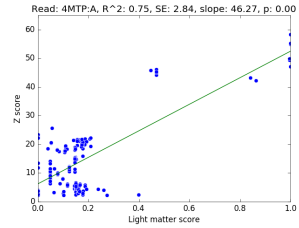


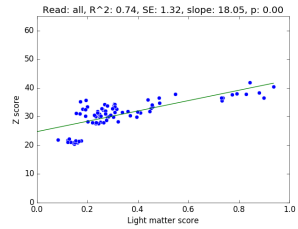


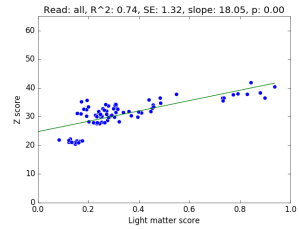


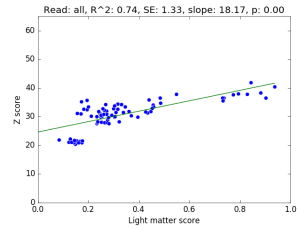


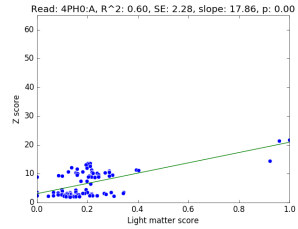


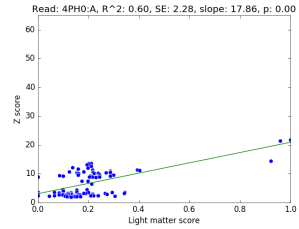


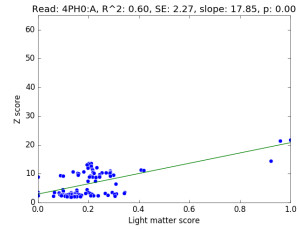


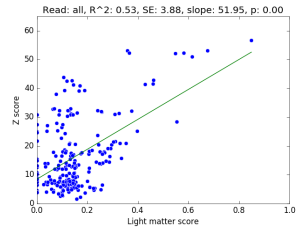


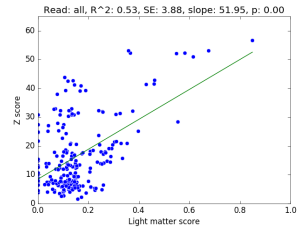


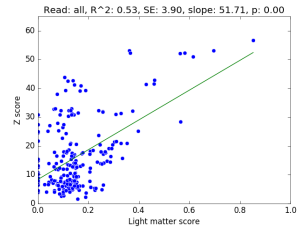












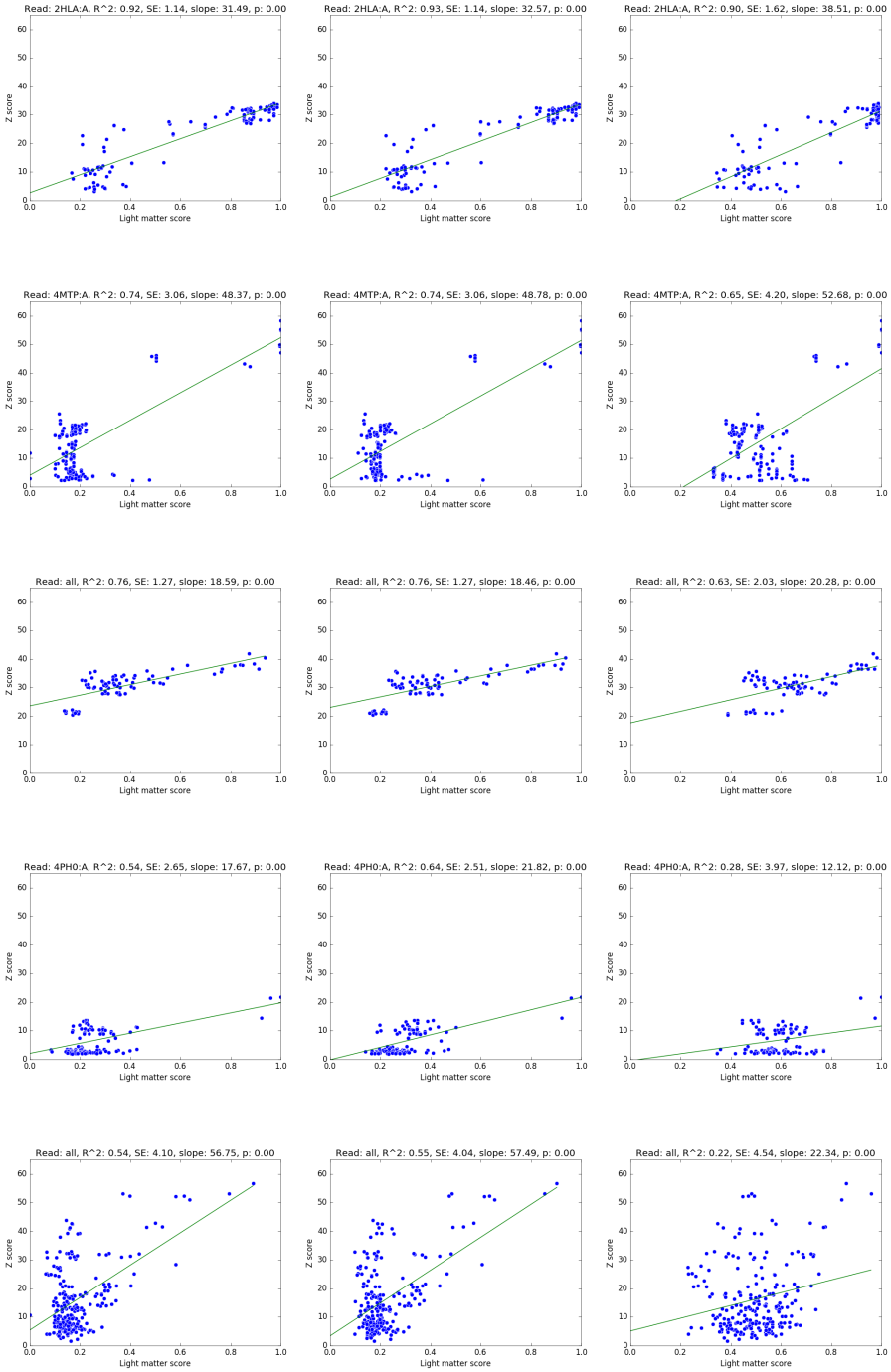

Read: 2HLA:A, R^2: 0.92, SE: 1.14, slope: 31.49, p: 0.00
Read: 2HLA:A, R^2: 0.93, SE: 1.14, slope: 32.57, p: 0.00
Read: 2HLA:A, R^2: 0.90, SE: 1.62, slope: 38.51, p: 0.00
Read: 4MTP:A, R^2: 0.74, SE: 3.06, slope: 48.37, p: 0.00
Read: 4MTP:A, R^2: 0.74, SE: 3.06, slope: 48.78, p: 0.00
Read: 4MTP:A, R^2: 0.65, SE: 4.20, slope: 52.68, p: 0.00
Read: all, R^2: 0.76, SE: 1.27, slope: 18.59, p: 0.00
Read: all, R^2: 0.76, SE: 1.27, slope: 18.46, p: 0.00
Read: all, R^2: 0.63, SE: 2.03, slope: 20.28, p: 0.00
Read: 4PH0:A, R^2: 0.54, SE: 2.65, slope: 17.67, p: 0.00
Read: 4PH0:A, R^2: 0.64, SE: 2.51, slope: 21.82, p: 0.00
Read: 4PH0:A, R^2: 0.28, SE: 3.97, slope: 12.12, p: 0.00
Read: all, R^2: 0.54, SE: 4.10, slope: 56.75, p: 0.00
Read: all, R^2: 0.55, SE: 4.04, slope: 57.49, p: 0.00
Read: all, R^2: 0.22, SE: 4.54, slope: 22.34, p: 0.00
Z score
Light matter score

Figure C.5: Correlation of FeatureAAScore and Z-scores, using different values for the DistanceBase parameter. Rows, top to bottom: 1) 2HLA. 2) 4MTP. 3) HA. 4) 4PH0. 5) Polymerase. Columns, left to right: 1) DistanceBase: 1.01. 2) DistanceBase: 1.05. 3) DistanceBase: 1.1. 4) DistanceBase: 1.2. 5) DistanceBase: 1.35. 6) DistanceBase: 1.5. 'AC AlphaHelix_combined', 'AC ExtendedStrand', 'Prosite', 'EukaryoticLinearMotif', and 'AminoAcidsLm' landmark finders and 'Peak', 'Trough', and 'AminoAcids' trig point finders, a significance fraction of 0.01, a FeatureLengthBase of 1.01, DeltaScale of 1.0, and a MaxDistance of 1000 were used throughout. Lines indicate the linear regression, with coefficients in the title of each subfigure.

I tested four different values for the MaxDistance parameter, 10, 100, 500, and 1,000. The effect of different values for the MaxDistance parameter is minimal based on changes in the correlation coefficients of the linear regression (Fig. C.6). The light matter algorithm pairs features starting with the landmark closest to the landmark from which the pairing starts, and going to the next closest and so on, and subsequently from closest to furthest trig point. There is also a limit on the number of pairs that a landmark can be part of, which I left at 10 throughout, as a higher number of pairs per landmark negatively impacts the speed of the algorithm (not shown). Thus, if 10 features are within 100 amino acids of a landmark, we expect to only detect a difference between a MaxDistance of 10 (column one) and a MaxDistance of 100 (column two), which is what we see in Fig. C.6. A larger MaxDistance allows for more distant features to form a pair. In the context of matching short reads, it makes sense to make the MaxDistance parameter to be at most as long as the length of the read, hence allowing all features in a short sequence to pair. Should an insertion have taken place, a bigger MaxDistance parameter would also allow the pairing of features on both sides of the insertion, and the sequence could then still be compared to a reference without the insertion.

The analyses of different parameter combinations allows the following conclusions: landmarks and trig points should be used, with the best combination of features being the 'AC AlphaHelix_combined', 'AC ExtendedStrand', 'Prosite', 'EukaryoticLinearMotif', and 'AminoAcidsLm' landmarks, and the 'Peak', 'Trough', and 'AminoAcids' trig points. The FeatureLengthBase and MaxDistance parameters have limited effect on the performance of the light matter algorithm. A DistanceBase of 1.35 results in the best correlation between light matter scores and Dali Z-scores. The results of similar analyses presented in chapter 8 have already shown that even when secondary structures are identified correctly, the matching and scoring as currently implemented in the light matter algorithm does not achieve acceptable results. This is because there is often a wide range of light matter scores that are assigned to a narrow range of Z-scores and vice versa, as well as high light matter scores assigned to matches with low Z-scores and vice versa. The results in this section also show that even when using additional finders, and a variety of parameters, these problems persist.

C. LIGHT MATTER ALGORITHM: PRELIMINARY RESULTS

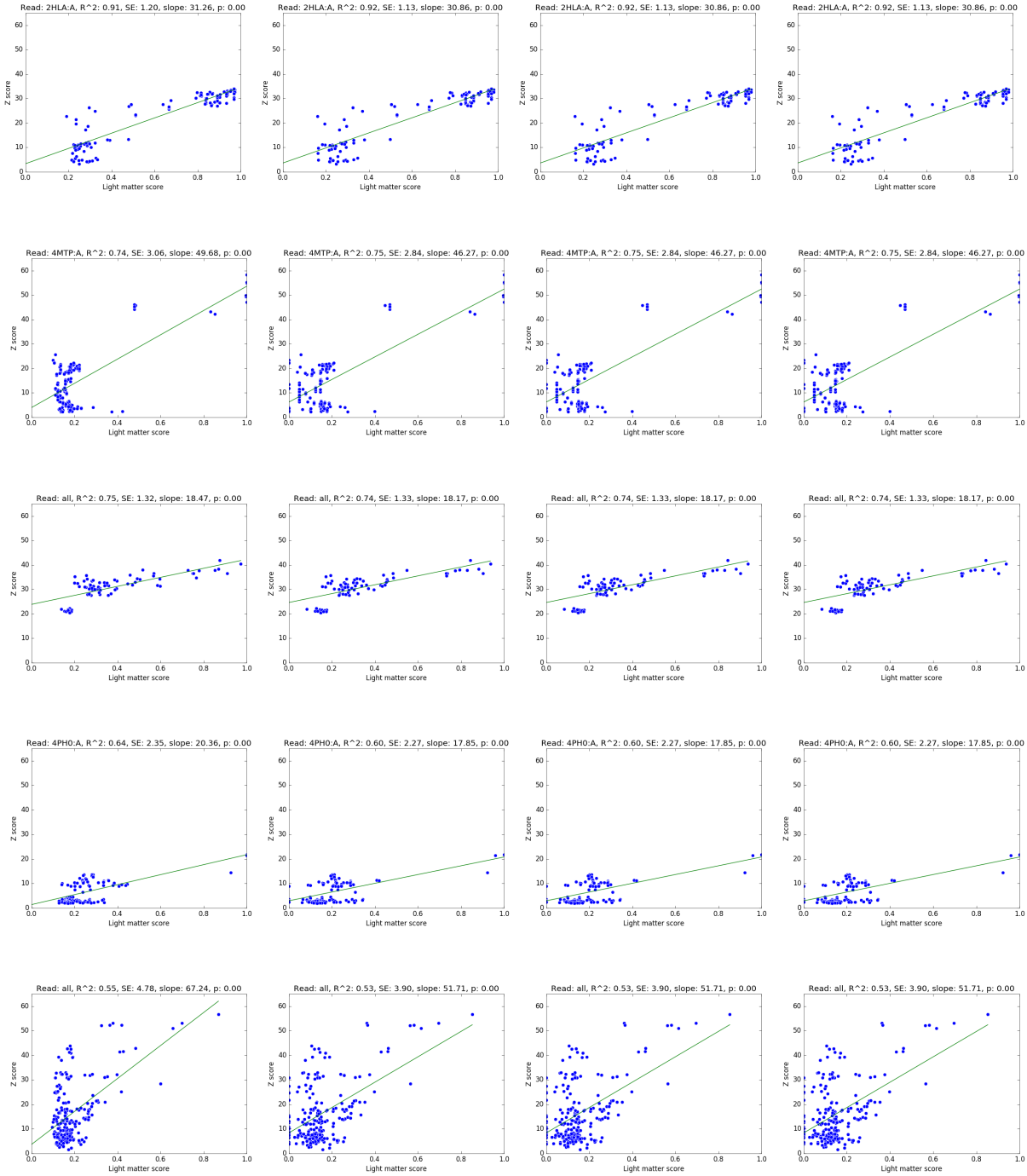


Figure C.6: Correlation of FeatureAAScore and Z-scores, using different values for the MaxDistance parameter. Rows, top to bottom: 1) 2HLA. 2) 4MTP. 3) HA. 4) 4PH0. 5) Polymerase. Columns, left to right: 1) MaxDistance: 10. 2) MaxDistance: 100. 3) MaxDistance: 500. 4) MaxDistance: 1000. 'AC AlphaHelix_combined', 'AC ExtendedStrand', 'Prosite', 'EukaryoticLinearMotif', and 'AminoAcidsLm' landmark finders and 'Peak', 'Trough', and 'AminoAcids' trig point finders, and a significance fraction of 0.01, a FeatureLengthBase of 1.01, DistanceBase of 1.1, and DeltaScale of 1 were used throughout. Lines indicate the linear regression, with coefficients in the title of each subfigure.

C.2 INVESTIGATING LIGHT MATTER SCORING IRREGULARITIES BY EXAMINING THREE-DIMENSIONAL PROTEIN STRUCTURES

The previous sections have shown that there are inconsistencies in the light matter scores assigned to matches by the FeatureAAScore scoring method, in comparison to their associated Z-scores. Namely, similar light matter scores are assigned to matches with a wide range of Z-scores, matches with similar Z-scores are assigned a range of light matter scores, and matches with low Z-scores are assigned high light matter scores. While broad insights can be gained by looking at overall correlations between light matter scores and Dali Z-scores, it is useful to investigate some matches in isolation, to better understand why the scores may be higher or lower than expected. The work presented in this section relies on visualisations of how matching and non-matching features are distributed on the three-dimensional protein structures, to understand of how feature identification and distribution influences the light matter scores computed for a match. Using the 4MTP and Polymerase test datasets, I will show examples for the following scenarios:

- Matches with similar light matter scores but Z-scores ranging from 2.2 to 25.6.
- Matches that have Z-scores between 32.2 and 33.0, but light matter scores ranging from 0.2 to 0.6.
- Matches with Z-scores lower than eight, but light matter scores ranging from 0.45 to 0.61.

For the 4MTP dataset, the 'AC AlphaHelix_combined', and 'AC ExtendedStrand' landmark finders were used, with no trig points. For the Polymerase dataset, the 'AC AlphaHelix_combined', 'AC ExtendedStrand', and 'AminoAcidsLm' landmark finders were used, with no trig points. The other parameters were:

```
ScoringMethod=FeatureAAScore,
SignificanceMethod=HashFraction,
SignificanceCutoff=0.01, FeatureLenghtBase=1.0,
MaxDistance=10000, LimitPerLandmark=10,
DistanceBase=1.1, DeltaScale=1.0.
```

C.2.1 Matches with similar light matter scores but Z-scores ranging from 2.2 to 25.6

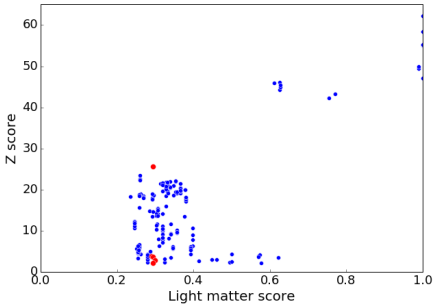


Figure C.7: Scatter plot of FeatureAAScore light matter scores against Dali Z-scores for the 4MTP test dataset. Matches with a light matter score between 0.29 and 0.31 are shown in red.

Figure C.7 shows a scatter plot of the FeatureAAScore light matter scores against the Dali Z-scores computed for the 4MTP test dataset using the parameters described above. The plot shows a cluster of matches at the bottom to the left of the figure, with light matter scores between 0.2 and 0.4, and Z-scores between 2.2 and 25.6. This suggests that the light matter scoring does not adequately separate between proteins with different structural similarity, as measured by the Z-score. I highlight twelve matches that have a light matter score between 0.29 and 0.31, and all but one have a Z-score between 2.2 and 3.8, and an outlier having a Z-score of 25.6 (plotted as red dots, in Fig. C.7). Examining the structures visually shows that in the comparisons with low Z-scores, one of the structures being compared has a molecule bound to it, possibly distorting its shape (an example of which is shown in Fig. C.8), whereas this is not the case in the match with the high Z-score (Fig. C.9). However, the match with the high Z-score is assigned a light matter score of 0.3, lower than expected given its high Z-score. Visually comparing the three-dimensional structures (Fig. C.9), shows that even though the structures look similar visually, only few features are identified by the 'AC AlphaHelix_combined' and 'AC ExtendedStrand' feature finders. Furthermore, the features that are matching do not belong to secondary structures that are structurally equivalent, meaning that for example, a pair containing an alpha helix matches another pair containing an alpha helix, but the two alpha helices are from different regions of the structure. The light matter score thus does not adequately describe the structural similarity between the two structures in Fig. C.9.

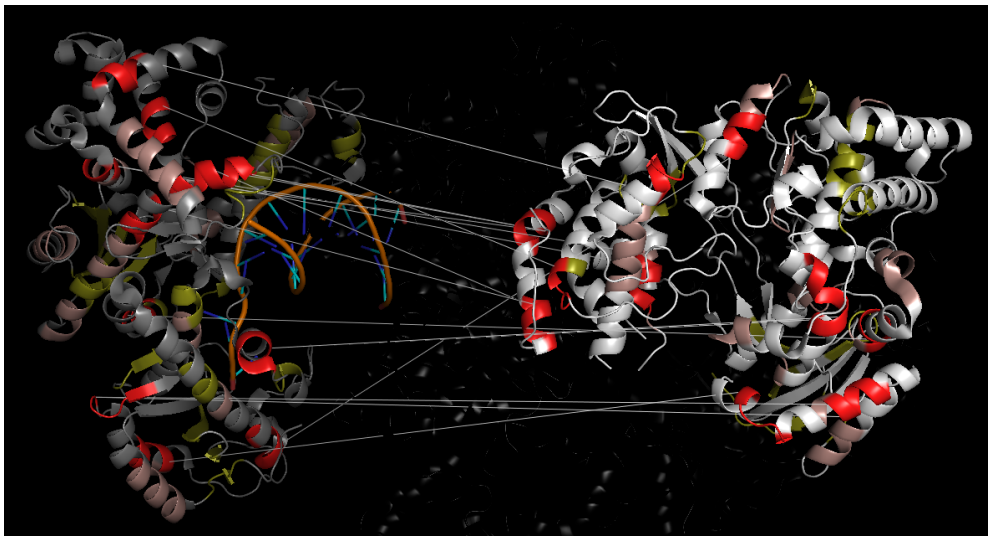


Figure C.8: Structures of 4MTP (right) and 1KLN (left). 'AC AlphaHelix_combined' in brown, 'AC ExtendedStrand' in green, matching features in red. White lines show features in matching pairs. FeatureAAScore light matter score: 0.3. Z-score: 2.2.

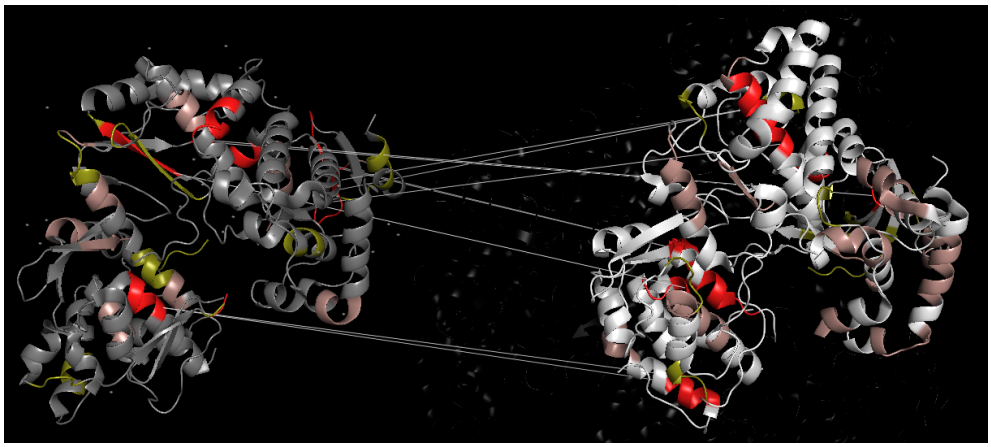


Figure C.9: Structures of 4MTP (right) and 2CJQ (left). 'AC AlphaHelix_combined' in brown, 'AC ExtendedStrand' in green, matching features in red. White lines show features in matching pairs. FeatureAAScore light matter score: 0.3. Z-score: 25.6.

C.2.2 Matches that have similar Z-scores (32.2 to 33.0) but light matter scores ranging from 0.21 to 0.55

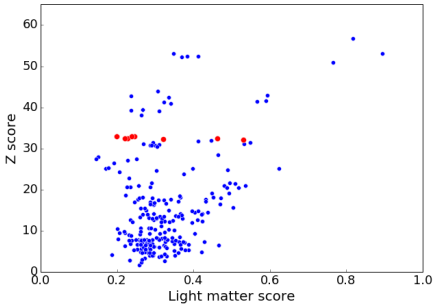


Figure C.10: Scatter plot of FeatureAAScore light matter scores against Dali Z-scores for the Polymerase test dataset. Matches with a Z-score between 32.2 and 33.0 are highlighted in red.

Figure C.10 shows a scatter plot of light matter and Z-scores computed for the Polymerase dataset, using the parameters outlined in the introduction to the current section. I am focussing on eight matches that have similar Z-scores between 32.2 and 33.0, but variable light matter scores, ranging from 0.21 to 0.55 (plotted as red dots in Fig. C.10). Visually inspecting the structures for the matches with the highest (0.55, Fig. C.11) and lowest (0.21, Fig. C.12) light matter score shows that the match that is assigned a high light matter score has a high number of secondary structures matching that are structurally equivalent (Fig. C.11). For the match with the worst light matter score (Fig C.12) on the other hand, matching features are not structurally equivalent. As in the previous section, the light matter score for the match in Fig. C.12 fails to reflect the structural similarity, due to inadequate feature identification, and because the pairs that take part in the score calculation are not structurally equivalent.

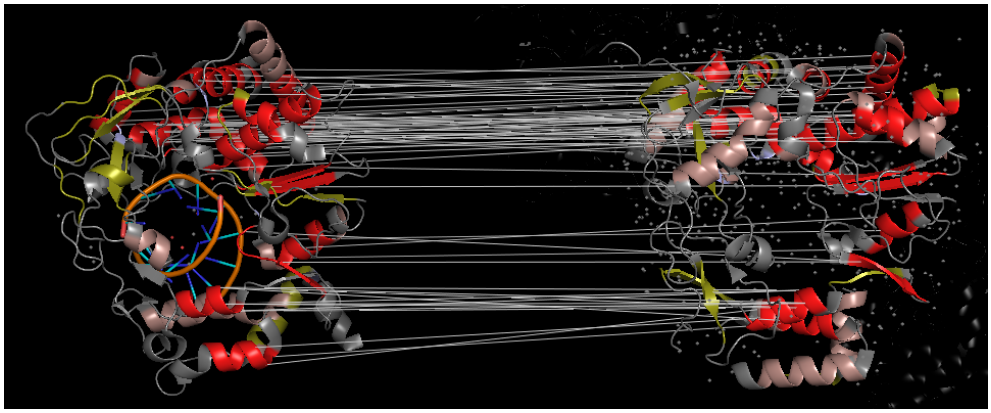


Figure C.11: Structures of 3UQS (right) and 2E9Z (left). 'AC AlphaHelix_combined' in brown, 'AC ExtendedStrand' in green, 'AminoAcidsLm' in pink, matching features in red. White line show features in matching pairs. FeatureAAScore light matter score: 0.55. Z-score: 32.2.

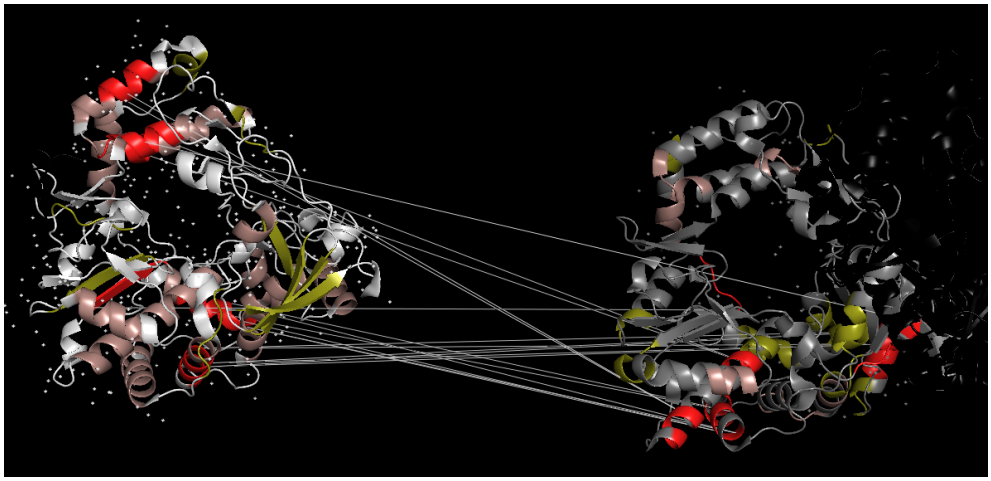


Figure C.12: Structures of 3CDW (right) and 1KHV (left). 'AC AlphaHelix_combined' in brown, 'AC ExtendedStrand' in green, 'AminoAcidsLm' in pink, matching features in red. White line show features in matching pairs. FeatureAAScore light matter score: 0.21. Z-score: 33.0.

C.2.3 Matches with low Z-scores and high light matter scores

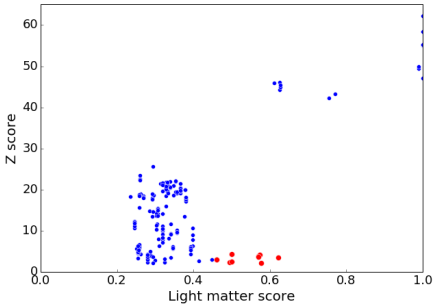


Figure C.13: Scatter plot of FeatureAAScore light matter scores against Dali Z-scores for the 4MTP test dataset. Matches with a light matter score between 0.45 and 0.61 and a Z-score below 8 are shown in red.

Figure C.13 shows a scatter plot of light matter scores and Z-scores from the 4MTP dataset. I would like to focus on two out of nine matches with light matter scores between 0.45 and 0.61 and with Z-scores below 8.0 highlighted in red. Why does the light matter algorithm assign light matter scores that are seemingly too high? In the two matches with the highest and the lowest light matter scores, the sequences coding for these protein structures are of different lengths (634 versus 308 amino acids for 4MTP and 2G1H, and 634 versus 98 amino acids for 4MTP and 1H6K), and the resulting structures are dis-similar (Figs. C.14 and C.15). The visual impression thus agrees with the low Z-scores. The location of the matching features (as indicated by white lines) also shows that the matching features are not structurally equivalent, leading to artificially inflated light matter scores. Furthermore, the length discrepancy between the sequences that code for the structures, can also artificially increase the score, due to the length normalisation that is performed in the FeatureAAScore calculation, which rewards matches where the matching pairs are distributed across the entire query and / or subject.

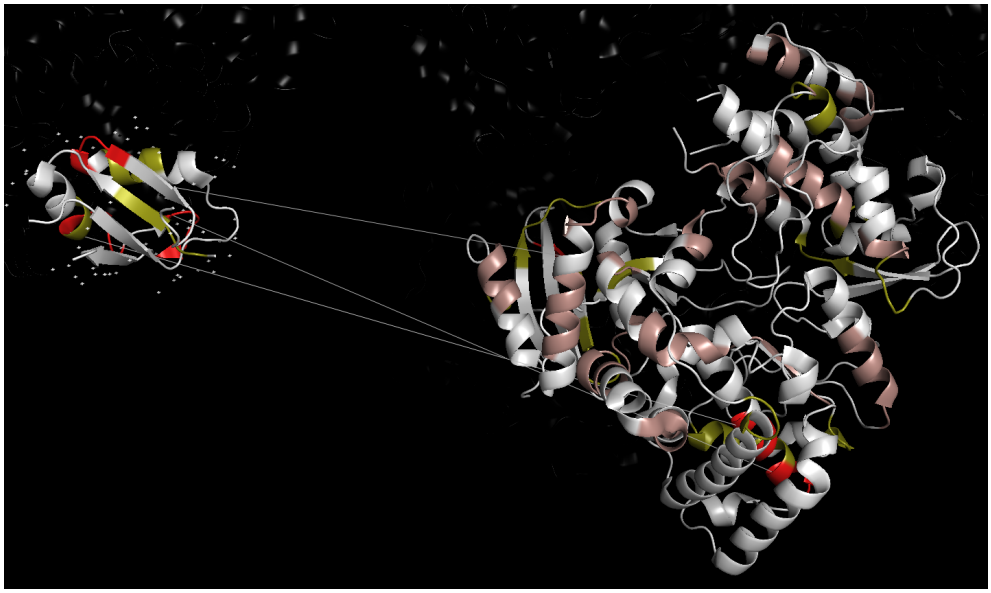


Figure C.14: Structures of 4MTP (right) and 1H6K (right). 'AC AlphaHelix_combined' in brown, 'AC ExtendedStrand' in green, 'AminoAcidsLm' in pink, matching features in red. White lines show features in matching pairs. FeatureAAScore light matter score: 0.61. Z-score: 3.5.

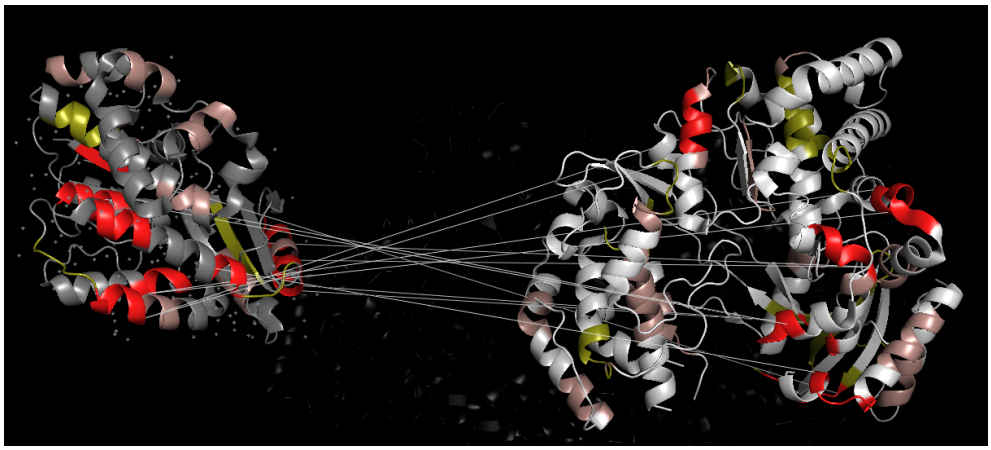


Figure C.15: Structures of 4MTP (left) and 2G1H (right). 'AC AlphaHelix_combined' in brown, 'AC ExtendedStrand' in green, 'AminoAcidsLm' in pink, matching features in red. White lines show features in matching pairs. FeatureAAScore light matter score: 0.47. Z-score: 3.0.

C.2.4 Conclusion

I found three reasons for irregularities in the correspondence between light matter and Z-scores. The first is related to differences in protein structure, caused by ligands, leading to Z-scores that are lower than expected relative to the light matter score. The second is that low light matter scores can be caused when the finders do not identify adequate features. The low light matter scores that are observed from such matches often stem from matches between features that are not structurally equivalent. And finally, light matter scores may be artificially inflated, either because a short sequence is compared to a long one (an artefact of the FeatureAAScore calculation), or again because of matches between features that are not structurally equivalent. While these conclusions are based on a very small number of sequence and structure comparisons, they nevertheless highlight two areas that need to be improved. The first is the FeatureAAScore calculation, to improve the scoring of matches between sequences of different lengths. The second is the feature identification, which needs to ensure that features are identified correctly and consistently (e.g., an alpha helix is always identified as an alpha helix), and also that structurally equivalent features match, and matches between features that are not structurally equivalent do not dominate the score calculation.

BIBLIOGRAPHY

[1] E V Koonin and V V Dolja. A virocentric perspective on the evolution of life. *Current Opinion in Virology*, 3(5):546–557, 2013.

[2] B N Fields, D M Knipe, and P M Howley. Fields Virology, 6th Edition, 2013.

[3] M A O'Malley. The ecological virus. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 59:71–79, 2016.

[4] F Rohwer and R V Thurber. Viruses manipulate the marine environment. *Nature*, 459(7244):207, 2009.

[5] M Worobey, A Bjork, and J O Wertheim. Point, counterpoint: the evolution of pathogenic viruses and their human hosts. *Annual Review of Ecology, Evolution, and Systematics*, 38:515–540, 2007.

[6] J M Alves, M Carneiro, J Y Cheng, A L de Matos, M M Rahman, L Loog, P F Campos, N Wales, A Eriksson, A Manica, et al. Parallel adaptation of rabbit populations to myxoma virus. *Science*, 363(6433):1319–1326, 2019.

[7] G L Smith, C T O Benfield, C Maluquer de Motes, M Mazzon, S W J Ember, B J Ferguson, and R P Sumner. Vaccinia virus immune evasion: mechanisms, virulence and immunogenicity. *Journal of General Virology*, 94(Pt 11):2367–2392, 2013.

[8] S Mi, X Lee, X Li, G M Veldman, H Finnerty, L Racie, E Lavallie, X-Y Tang, P Edouard, S Howes, et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, 403(6771):785, 2000.

[9] J H Kuhn, L E Dodd, V Wahl-Jensen, S R Radoshitzky, S Bavari, and P B Jahrling. Evaluation of perceived threat differences posed by filovirus variants. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 9(4):361–371, 2011.

[10] F Fenner, D A Henderson, I Arita, S Jezek, I D Ladnyi, et al. *Smallpox and its eradication*, volume 6. World Health Organization Geneva, 1988.

[11] WHO. Rabies Fact Sheet. `https://www.who.int/en/news-room/fact-sheets/detail/rabies`, 2019. Accessed: 2019-03-30.

[12] The Center for Food Security and Public Health, Iowa State University. Avian Influenza. `http://www.cfsph.iastate.edu/Factsheets/pdfs/highly_pathogenic_avian_influenza-citations.pdf`, 2015. Accessed: 2019-03-30.

[13] DJ Alexander, G Parsons, and RJ Manvell. Experimental assessment of the pathogenicity of eight avian influenza A viruses of H5 subtype for chickens, turkeys, ducks and quail. *Avian Pathology*, 15(4):647–662, 1986.

[14] José Manuel Sánchez-Vizcaíno, Alberto Laddomada, and Marisa L Arias. African swine fever virus. *Diseases of Swine*, pages 443–452, 2019.

[15] K E Jones, N G Patel, M A Levy, A Storeygard, D Balk, J L Gittleman, and P Daszak. Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993, 2008.

[16] World population prospects: The 2017 revision, key findings and advance tables. Technical Report ESA/P/WP/248, United Nations, Department of Economic and Social Affairs, Population Division, 2017.

[17] K G Austin, M González-Roglich, D Schaffer-Smith, A M Schwantes, and J J Swenson. Trends in size of tropical deforestation events signal increasing dominance of industrial-scale drivers. *Environmental Research Letters*, 12(5):054009, 2017.

[18] World urbanization prospects: The 2014 revision, highlights. Technical Report ST/ESA/SER.A/352, United Nations, Department of Economic and Social Affairs, Population Division, 2014.

[19] R A Weiss and A J McMichael. Social and environmental risk factors in the emergence of infectious diseases. *Nature Medicine*, 10(12s):S70, 2004.

[20] J L Mokili, F Rohwer, and B E Dutilh. Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology*, 2(1):63–77, 2012.

[21] D S Leland and C C Ginocchio. Role of cell culture for virus detection in the age of technology. *Clinical Microbiology Reviews*, 20(1):49–78, 2007.

[22] R I Amann, W Ludwig, and K-H Schleifer. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews*, 59(1):143–169, 1995.

[23] National Human Genome Research Institute (NHGRI). DNA Sequencing Costs: Data. `https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data`, 2018. Accessed: 2019-04-28.

[24] K Chen and L Pachter. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Computational Biology*, 1(2):e24, 2005.

[25] C Camacho, G Coulouris, V Avagyan, N Ma, J Papadopoulos, K Bealer, and T L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

[26] S Canzar and S L Salzberg. Short read mapping: An algorithmic tour. *Proceedings of the IEEE*, 105(3):436–458, 2017.

[27] L Fancello, D Raoult, and C Desnues. Computational tools for viral metagenomics and their application in clinical research. *Virology*, 434(2):162–174, 2012.

[28] A Nasir, P Forterre, K M Kim, and G Caetano-Anollés. The distribution and impact of viral lineages in domains of life. *Frontiers in Microbiology*, 5:194, 2014.

[29] Y-Z Zhang, M Shi, and E C Holmes. Using metagenomics to characterize an expanding virosphere. *Cell*, 172(6):1168–1172, 2018.

[30] Jemma L Geoghegan and Edward C Holmes. Predicting virus emergence amid evolutionary noise. *Open Biology*, 7(10):170189, 2017.

[31] C Desnues, B Rodriguez-Brito, S Rayhawk, S Kelley, T Tran, M Haynes, H Liu, M Furlan, L Wegley, B Chau, et al. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature*, 452(7185):340, 2008.

[32] K Illergård, D H Ardell, and A Elofsson. Structure is three to ten times more conserved than sequence - a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499–508, 2009.

[33] B F Koel, D F Burke, T M Bestebroer, S van der Vliet, G C M Zondag, G Vervaet, E Skepner, N S Lewis, M I J Spronken, C A Russell, et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*, 342(6161):976–979, 2013.

[34] K Tsangaras and A D Greenwood. *Paleovirology: Viral Sequences from Historical and Ancient DNA*, pages 139–162. Springer International Publishing, Cham, 2018.

[35] S Duffy, L A Shackelton, and E C Holmes. Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics*, 9(4):267, 2008.

[36] D Paraskevis, K Angelis, G Magiorkinis, E Kostaki, S Y W Ho, and A Hatzakis. Dating the origin of hepatitis B virus reveals higher substitution rate and adaptation on the branch leading to F/H genotypes. *Molecular Phylogenetics and Evolution*, 93:44–54, 2015.

[37] Y Zhou and E C Holmes. Bayesian estimates of the evolutionary rate and age of hepatitis B virus. *Journal of Molecular Evolution*, 65(2):197–205, 2007.

[38] P Aiewsakun and A Katzourakis. Time dependency of foamy virus evolutionary rate estimates. *BMC Evolutionary Biology*, 15(1):119, 2015.

[39] S Pääbo, H Poinar, D Serre, V Jaenicke-Després, J Hebler, N Rohland, M Kuch, J Krause, L Vigilant, and M Hofreiter. Genetic analyses from ancient DNA. *Annual Review of Genetics*, 38:645–679, 2004.

[40] R Nielsen, J M Akey, M Jakobsson, J K Pritchard, S Tishkoff, and E Willerslev. Tracing the peopling of the world through genomics. *Nature*, 541(7637):302–310, 2017.

[41] M E Allentoft, M Sikora, K-G Sjögren, S Rasmussen, M Rasmussen, J Stenderup, P Damgaard, H Schroeder, T Ahlström, L Vinner, et al. Population genomics of Bronze Age Eurasia. *Nature*, 522(7555):167–172, 2015.

[42] W Haak, I Lazaridis, N Patterson, N Rohland, S Mallick, B Llamas, G Brandt, S Nordenfelt, E Harney, K Stewardson, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207, 2015.

[43] P de Barros Damgaard, R Martiniano, J Kamm, J V Moreno-Mayar, G Kroonen, M Peyrot, G Barjamovic, S Rasmussen, C Zacho, N Baimukhanov, et al. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science*, 360(6396):eaar7711, 2018.

[44] P de Barros Damgaard, N Marchi, S Rasmussen, M Peyrot, G Renaud, T Korneliussen, J V Moreno-Mayar, M W Pedersen, A Goldberg, E Usmanova, et al. 137 ancient human genomes from across the Eurasian steppes. *Nature*, 557(7705):369, 2018.

[45] P Skoglund, C Posth, K Sirak, M Spriggs, F Valentin, S Bedford, G R Clark, C Reepmeyer, F Petchey, D Fernandes, et al. Genomic insights into the peopling of the Southwest Pacific. *Nature*, 538(7626):510, 2016.

[46] M Raghavan, P Skoglund, K E Graf, M Metspalu, A Albrechtsen, I Moltke, S Rasmussen, T W Stafford Jr, L Orlando, E Metspalu, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, 505(7481):87, 2014.

[47] J V Moreno-Mayar, L Vinner, P de Barros Damgaard, C de la Fuente, J Chan, J P Spence, M E Allentoft, T Vimala, F Racimo, T Pinotti, et al. Early human dispersals within the Americas. *Science*, 362(6419):eaav2621, 2018.

[48] C Posth, N Nakatsuka, I Lazaridis, P Skoglund, S Mallick, T C Lamnidis, N Rohland, K Nägele, N Adamski, E Bertolini, et al. Reconstructing the deep population history of Central and South America. *Cell*, 175(5):1185–1197, 2018.

[49] K Prüfer, F Racimo, N Patterson, F Jay, S Sankararaman, S Sawyer, A Heinze, G Renaud, P H Sudmant, C De Filippo, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43, 2014.

[50] D Reich, R E Green, M Kircher, J Krause, N Patterson, E Y Durand, B Viola, A W Briggs, U Stenzel, P L F Johnson, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–1060, 2010.

[51] S Fan, M E B Hansen, Y Lo, and S A Tishkoff. Going global by adapting local: A review of recent human adaptation. *Science*, 354(6308):54–59, 2016.

[52] C M Beall. Andean, Tibetan, and Ethiopian patterns of adaptation to high-altitude hypoxia. *Integrative and Comparative Biology*, 46(1):18–24, 2006.

[53] T Bersaglieri, P C Sabeti, N Patterson, T Vanderploeg, S F Schaffner, J A Drake, M Rhodes, D E Reich, and J N Hirschhorn. Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6):1111–1120, 2004.

[54] Matteo Fumagalli, Ida Moltke, Niels Grarup, Fernando Racimo, Peter Bjerregaard, Marit E Jørgensen, Thorfinn S Korneliussen, Pascale Gerbault, Line Skotte, Allan Linneberg, et al. Greenlandic inuit show genetic signatures of diet and climate adaptation. *Science*, 349(6254):1343–1347, 2015.

[55] M A Ilardo, I Moltke, T S Korneliussen, J Cheng, A J Stern, F Racimo, P de Barros Damgaard, M Sikora, A Seguin-Orlando, S Rasmussen, et al. Physiological and genetic adaptations to diving in sea nomads. *Cell*, 173(3):569–580, 2018.

[56] R Higuchi, B Bowman, M Freiberger, O A Ryder, and A C Wilson. DNA sequences from the quagga, an extinct member of the horse family. *Nature*, 312(5991):282, 1984.

[57] E Cappellini, A Prohaska, F Racimo, F Welker, M W Pedersen, M E Allentoft, P de Barros Damgaard, P Gutenbrunner, J Dunne, S Hammann, M Roffet-Salque, M Ilardo, J V Moreno-Mayar, Y Wang, M Sikora, L Vinner, J Cox, R P Evershed, and E Willerslev. Ancient Biomolecules and Evolutionary Inference. *Annual Review of Biochemistry*, 87:1029–1060, 2018.

[58] M Rasmussen, Y Li, S Lindgreen, J S Pedersen, A Albrechtsen, I Moltke, M Metspalu, E Metspalu, T Kivisild, R Gupta, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 463(7282):757, 2010.

[59] R E Green, J Krause, A W Briggs, T Maricic, U Stenzel, M Kircher, N Patterson, H Li, W Zhai, M H-Y Fritz, et al. A draft sequence of the Neandertal genome. *Science*, 328(5979):710–722, 2010.

[60] P Skoglund, H Malmström, M Raghavan, J Storå, P Hall, W Willerslev, M T P Gilbert, A Götherström, and M Jakobsson. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*, 336(6080):466–469, 2012.

[61] E Willerslev, E Cappellini, W Boomsma, R Nielsen, M B Hebsgaard, T B Brand, M Hofreiter, M Bunce, H N Poinar, D Dahl-Jensen, et al. Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science*, 317(5834):111–114, 2007.

[62] L Orlando, A Ginolhac, G Zhang, D Froese, A Albrechtsen, M Stiller, M Schubert, E Cappellini, B Petersen, I Moltke, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, 499(7456):74, 2013.

[63] E Willerslev and A Cooper. Review paper. Ancient DNA. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1558):3–16, 2005.

[64] T Lindahl. Instability and decay of the primary structure of DNA. *Nature*, 362(6422):709, 1993.

[65] S Pääbo. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences*, 86(6):1939–1943, 1989.

[66] M Höss, P Jaruga, T H Zastawny, M Dizdaroglu, and S Pääbo. DNA damage and DNA sequence retrieval from ancient tissues. *Nucleic Acids Research*, 24(7):1304–1307, 1996.

[67] L Orlando, M T P Gilbert, and E Willerslev. Reconstructing ancient genomes and epigenomes. *Nature Reviews Genetics*, 16(7):395–408, 2015.

[68] M E Allentoft, M Collins, D Harker, J Haile, C L Oskam, M L Hale, P F Campos, J A Samaniego, M T P Gilbert, E Willerslev, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1748):4724–4733, 2012.

[69] L Kistler, R Ware, O Smith, M Collins, and R G Allaby. A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic Acids Research*, 45(11):6310–6320, 2017.

[70] S A Fish, T J Shepherd, T J McGenity, and W D Grant. Recovery of 16S ribosomal RNA gene fragments from ancient halite. *Nature*, 417(6887):432–436, 2002.

[71] S R Woodward, N J Weyand, and M Bunnell. DNA sequence from Cretaceous period bone fragments. *Science*, 266(5188):1229–1232, 1994.

[72] S Pääbo. Molecular cloning of ancient Egyptian mummy DNA. *Nature*, 314(6012):644, 1985.

[73] J J Austin, A B Smith, and R H Thomas. Palaeontology in a molecular world: the search for authentic ancient DNA. *Trends in Ecology & Evolution*, 12(8):303–306, 1997.

[74] S B Hedges, M H Schweitzer, S Henikoff, M W Allard, D Young, Y Huyen, H Zischler, M Höss, O Handt, A von Haeseler, et al. Detecting dinosaur DNA. *Science*, 268(5214):1191–1194, 1995.

[75] M T P Gilbert, H-J Bandelt, M Hofreiter, and I Barnes. Assessing ancient DNA studies. *Trends in Ecology & Evolution*, 20(10):541–544, 2005.

[76] A Cooper and H N Poinar. Ancient DNA: do it right or not at all. *Science*, 289(5482):1139–1139, 2000.

[77] M Knapp, C Lalueza-Fox, and M Hofreiter. Re-inventing ancient human DNA. *Investigative Genetics*, 6(1):4, 2015.

[78] S Rasmussen, M E Allentoft, K Nielsen, L Orlando, M Sikora, K-G Sjögren, A G Pedersen, M Schubert, A Van Dam, C M O Kapel, et al. Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. *Cell*, 163(3):571–582, 2015.

[79] B Mühlemann, T C Jones, P de Barros Damgaard, M E Allentoft, I Shevnina, A Logvin, E Usmanova, I P Panyushkina, B Boldgiv, T Bazartseren, et al. Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature*, 557(7705):418, 2018.

[80] B Mühlemann, A Margaryan, P de Barros Damgaard, M E Allentoft, L Vinner, A J Hansen, A Weber, V I Bazaliiskii, M Molak, K Arneborg, et al. Ancient human parvovirus B19 in Eurasia reveals its long-term association with humans. *Proceedings of the National Academy of Sciences*, 115(29):7557–7562, 2018.

[81] L Orlando and A Cooper. Using ancient DNA to understand evolutionary and ecological processes. *Annual Review of Ecology, Evolution, and Systematics*, 45:573–598, 2014.

[82] D Reich. *Who we are and how we got here*, chapter Introduction, page xvii. Oxford University Press, 2018.

[83] P Skoglund and I Mathieson. Ancient genomics of modern humans: the first decade. *Annual Review of Genomics and Human Genetics*, 19:381–404, 2018.

[84] K G Daly, P M Delser, V E Mullin, A Scheu, V Mattiangeli, M D Teasdale, A J Hare, J Burger, M P Verdugo, M J Collins, et al. Ancient goat genomes reveal mosaic domestication in the Fertile Crescent. *Science*, 361(6397):85–88, 2018.

[85] P Skoglund, E Ersmark, E Palkopoulou, and L Dalén. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology*, 25(11):1515–1519, 2015.

[86] M W Pedersen, A Ruter, C Schweger, H Friebe, R A Staff, K K Kjeldsen, M L Z Mendoza, A B Beaudoin, C Zutter, N K Larsen, et al. Postglacial viability and colonization in North America's ice-free corridor. *Nature*, 537(7618):45, 2016.

[87] E Hagelberg, M Hofreiter, and C Keyser. Ancient DNA: the first three decades, 2015.

[88] F M Galassi, M E Habicht, and F J Rühli. Poliomyelitis in ancient Egypt? *Neurological Sciences*, 38(2):375–375, 2017.

[89] M J Papagrigorakis, C Yapijakis, and P N Synodinos. Paleomicrobiology: Past human infections. chapter 10, page 162. Springer, 2008.

[90] D J Ortner. *Identification of pathological conditions in human skeletal remains*. Academic Press, 2003.

[91] K Tsangaras and A D Greenwood. Museums and disease: using tissue archive and museum samples to study pathogens. *Annals of Anatomy-Anatomischer Anzeiger*, 194(1):58–73, 2012.

[92] M A Spyrou, K I Bos, A Herbig, and J Krause. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nature Reviews Genetics*, 20(6):323–340, 2019.

[93] C Warinner, A Herbig, A Mann, J A Fellows Yates, C L Weiß, H A Burbano, L Orlando, and J Krause. A robust framework for microbial archaeology. *Annual Review of Genomics and Human Genetics*, 18:321–356, 2017.

[94] S Marciniak and H N Poinar. Ancient pathogens through human history: a paleogenomic perspective. In *Paleogenomics*, pages 115–138. Springer, 2018.

[95] K I Bos, V J Schuenemann, G B Golding, H A Burbano, N Waglechner, B K Coombes, J B McPhee, S N DeWitte, M Meyer, S Schmedes, et al. A draft genome of Yersinia pestis from victims of the Black Death. *Nature*, 478(7370):506–510, 2011.

[96] M Feldman, M Harbeck, M Keller, M A Spyrou, A Rott, B Trautmann, H C Scholz, B Päffgen, J Peters, M McCormick, et al. A high-coverage Yersinia pestis Genome from a 6th-century Justinianic Plague Victim. *Molecular Biology and Evolution*, page msw170, 2016.

[97] K I Bos, A Herbig, J Sahl, N Waglechner, M Fourment, S A Forrest, J Klunk, V J Schuenemann, D Poinar, M Kuch, et al. Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. *elife*, 5:e12994, 2016.

[98] N Rascovan, K-G Sjögren, K Kristiansen, R Nielsen, E Willerslev, C Desnues, and S Rasmussen. Emergence and spread of basal lineages of yersinia pestis during the Neolithic decline. *Cell*, 176(1-2):295–305, 2019.

[99] F Maixner, B Krause-Kyora, D Turaev, A Herbig, M R Hoopmann, J L Hallows, U Kusebauch, E E Vigl, P Malfertheiner, F Megraud, et al. The 5300-year-old Helicobacter pylori genome of the Iceman. *Science*, 351(6269):162–165, 2016.

[100] K I Bos, K M Harkins, A Herbig, M Coscolla, N Weber, I Comas, S A Forrest, J M Bryant, S R Harris, V J Schuenemann, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, 514(7523):494–497, 2014.

[101] V J Schuenemann, P Singh, T A Mendum, B Krause-Kyora, G Jäger, K I Bos, A Herbig, C Economou, A Benjak, P Busso, et al. Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science*, 341(6142):179–183, 2013.

[102] V J Schuenemann, C Avanzi, B Krause-Kyora, A Seitz, A Herbig, S Inskip, M Bonazzi, W Reiter, C Urban, D D Pedersen, et al. Ancient genomes reveal a high diversity of Mycobacterium leprae in medieval Europe. *PLoS Pathogens*, 14(5):e1006997, 2018.

[103] G L Kay, M J Sergeant, V Giuffra, P Bandiera, M Milanese, B Bramanti, R Bianucci, and M J Pallen. Recovery of a medieval Brucella melitensis genome using shotgun metagenomics. *MBio*, 5(4):e01337–14, 2014.

[104] V J Schuenemann, A K Lankapalli, R Barquera, E A Nelson, D I Hernández, V A Alonzo, K I Bos, L M Morfín, A Herbig, and J Krause. Historic Treponema pallidum genomes from Colonial Mexico retrieved from archaeological remains. *PLoS Neglected Tropical Diseases*, 12(6):e0006447, 2018.

[105] A J Vågene, A Herbig, M G Campana, N M R García, C Warinner, S Sabin, M A Spyrou, A A Valtueña, D Huson, N Tuross, et al. Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nature Ecology & Evolution*, 2(3):520–528, 2018.

[106] S Marciniak, T L Prowse, D A Herring, J Klunk, M Kuch, A T Duggan, L Bondioli, E C Holmes, and H N Poinar. Plasmodium falciparum malaria in 1st–2nd century CE southern italy. *Current Biology*, 26(23):R1220–R1222, 2016.

[107] C Warinner, J F M Rodrigues, R Vyas, C Trachsel, N Shved, J Grossmann, A Radini, Y Hancock, R Y Tito, S Fiddyment, et al. Pathogens and host immunity in the ancient human oral cavity. *Nature Genetics*, 46(4):336–344, 2014.

[108] A M Devault, T D Mortimer, A Kitchen, H Kiesewetter, J M Enk, G B Golding, J Southon, M Kuch, A T Duggan, W Aylward, et al. A molecular portrait of maternal sepsis from Byzantine Troy. *elife*, 6:e20983, 2017.

[109] J K Taubenberger, A H Reid, A E Krafft, K E Bijwaard, and T G Fanning. Initial genetic characterization of the 1918 "spanish" influenza virus. *Science*, 275(5307):1793–1796, 1997.

[110] M Worobey. Phylogenetic evidence against evolutionary stasis and natural abiotic reservoirs of influenza A virus. *Journal of Virology*, 82(7):3769–3774, 2008.

[111] M Peyambari, S Warner, N Stoler, D Rainer, and M J Roossinck. A 1000 year-old RNA virus. *Journal of Virology*, 93(1):e01188–18, 2018.

[112] B Krause-Kyora, J Susat, F M Key, D Kühnert, E Bosse, A Immel, C Rinne, S-C Kornell, D Yepes, S Franzenburg, et al. Neolithic and medieval virus genomes reveal complex evolution of hepatitis B. *eLife*, 7:e36666, 2018.

[113] Z Patterson Ross, J Klunk, G Fornaciari, V Giuffra, S Duchêne, A T Duggan, D Poinar, M W Douglas, J-S Eden, E C Holmes, et al. The paradox of HBV evolution as revealed from a 16th century mummy. *PLoS Pathogens*, 14(1):e1006750, 2018.

[114] K Bar-Gal, M J Kim, A Klein, D H Shin, C S Oh, J W Kim, T-H Kim, S B Kim, P R Grant, O Pappo, et al. Tracing hepatitis B virus to the 16th century in a Korean mummy. *Hepatology*, 56(5):1671–1680, 2012.

[115] P Pajer, J Dresler, H Kabíckova, L Písa, P Aganov, K Fucik, D Elleder, T Hron, V Kuzelka, P Velemínsky, et al. Characterization of two historic smallpox specimens from a Czech museum. *Viruses*, 9(8):200, 2017.

[116] A T Duggan, M F Perdomo, D Piombino-Mascali, S Marciniak, D Poinar, M V Emery, J P Buchmann, S Duchêne, R Jankauskas, M Humphreys, et al. 17th century variola virus reveals the recent history of smallpox. *Current Biology*, 26(24):3407–3412, 2016.

[117] P Biagini, C Thèves, P Balaresque, A Geraut, C Cannet, C Keyser, D Nikolaeva, P Gerard, S Duchesne, L Orlando, et al. Variola virus in a 300-year-old

Siberian mummy. *New England Journal of Medicine*, 367(21):2057–2059, 2012.

[118] E C Holmes. Freezing viruses in time. *Proceedings of the National Academy of Sciences*, 111(47):16643–16644, 2014.

[119] S Bédarida, O Dutour, A P Buzhilova, P De Micco, and P Biagini. Identification of viral DNA (Anelloviridae) in a 200-year-old dental pulp sample (Napoleon's Great Army, Kaliningrad, 1812). *Infection, Genetics and Evolution*, 11(2):358–362, 2011.

[120] P G Parker, E L Buckles, H Farrington, K Petren, N K Whiteman, R E Ricklefs, J L Bollmer, and G Jiménez-Uzcátegui. 110 years of Avipoxvirus in the Galapagos Islands. *PLoS One*, 6(1):e15989, 2011.

[121] O Smith, C Clapham, P Rose, Y Liu, J Wang, and R G Allaby. A complete ancient RNA genome: identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. *Scientific Reports*, 4:4003, 2014.

[122] C M Malmstrom, R Shu, E W Linton, L A Newton, and M A Cook. Barley yellow dwarf viruses (BYDVs) preserved in herbarium specimens illuminate historical disease ecology of invasive and native grasses. *Journal of Ecology*, 95(6):1153–1166, 2007.

[123] M S Tiee, R J Harrigan, H A Thomassen, and T B Smith. Ghosts of infections past: using archival samples to understand a century of monkeypox virus prevalence among host communities across space and time. *Royal Society open science*, 5(1):171089, 2018.

[124] G Fornaciari, K Zavaglia, L Giusti, C Vultaggio, and R Ciranni. Human papillomavirus in a 16th century mummy. *The Lancet*, 362(9390):1160, 2003.

[125] B B Larsen, K L Cole, and M Worobey. Ancient DNA provides evidence of 27,000-year-old papillomavirus infection and long-term codivergence with rodents. *Virus Evolution*, 4(1):vey014, 2018.

[126] M Toppinen, M F Perdomo, J U Palo, P Simmonds, S J Lycett, M Söderlund-Venermo, A Sajantila, and K Hedman. Bones hold the key to DNA virus history and epidemiology. *Scientific Reports*, 5:17226, 2015.

[127] M Legendre, J Bartoli, L Shmakova, S Jeudy, K Labadie, A Adrait, M Lescot, O Poirot, L Bertaux, C Bruley, et al. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proceedings of the National Academy of Sciences*, 111(11):4274–4279, 2014.

[128] S Calvignac, J-M Terme, S M Hensley, P Jalinot, A D Greenwood, and C Hänni. Ancient DNA identification of early 20th century simian T-cell leukemia virus type 1. *Molecular Biology and Evolution*, 25(6):1093–1098, 2008.

[129] H-C Li, T Fujiyoshi, H Lou, S Yashiki, S Sonoda, L Cartier, L Nunez, I Munoz, S Horai, and K Tajima. The presence of ancient human T-cell lymphotropic virus type I provirus DNA in an Andean mummy. *Nature Medicine*, 5(12):1428, 1999.

[130] J D Castello, S O Rogers, W T Starmer, C M Catranis, L Ma, C D Bachand, Y Zhao, and J E Smith. Detection of tomato mosaic tobamovirus RNA in ancient glacial ice. *Polar Biology*, 22(3):207–212, 1999.

[131] S Appelt, L Fancello, M Le Bailly, D Raoult, M Drancourt, and C Desnues. Viruses in a 14th-century coprolite. *Applied and Environmental Microbiology*, 80(9):2648–2655, 2014.

[132] T F F Ng, L-F Chen, Y Zhou, B Shapiro, M Stiller, P D Heintzman, A Varsani, N O Kondov, W Wong, X Deng, et al. Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. *Proceedings of the National Academy of Sciences*, 111(47):16842–16847, 2014.

[133] D J Smith, A S Lapedes, J C de Jong, T M Bestebroer, G F Rimmelzwaan, A D M E Osterhaus, and R A M Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, 2004.

[134] C A Russell, T C Jones, I G Barr, N J Cox, R J Garten, V Gregory, I D Gust, A W Hampson, A J Hay, A C Hurt, et al. The global circulation of seasonal influenza a (H3N2) viruses. *Science*, 320(5874):340–346, 2008.

[135] J M Fonville, S H Wilks, S L James, A Fox, M Ventresca, M Aban, L Xue, T C Jones, N M H Le, Q T Pham, et al. Antibody landscapes after influenza virus infection or vaccination. *Science*, 346(6212):996–1000, 2014.

[136] J K Pfeiffer and K Kirkegaard. Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. *PLoS Pathogens*, 1(2):e11, 2005.

[137] K M Pepin, S Lass, J R C Pulliam, A F Read, and J O Lloyd-Smith. Identifying genetic markers of adaptation for surveillance of viral host jumps. *Nature Reviews Microbiology*, 8(11):802, 2010.

[138] K M Peck and A S Lauring. Complexities of viral mutation rates. *Journal of Virology*, 92(14):e01031–17, 2018.

[139] B L Smith and C O Wilke. Virus Evolution: A new twist in measuring mutation rates. *eLife*, 6:e29586, 2017.

[140] L Glaser, J Stevens, D Zamarin, I A Wilson, A García-Sastre, T M Tumpey, C F Basler, J K Taubenberger, and P Palese. A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity. *Journal of Virology*, 79(17):11533–11536, 2005.

[141] E K Subbarao, W London, and B R Murphy. A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. *Journal of Virology*, 67(4):1761–1764, 1993.

[142] S Y W Ho and S Duchêne. Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology*, 23(24):5947–5965, 2014.

[143] E Zuckerkandl and L Pauling. Molecular disease, evolution and genetic heterogeneity. 1962.

[144] G M Jenkins, A Rambaut, O G Pybus, and E C Holmes. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *Journal of Molecular Evolution*, 54(2):156–165, 2002.

[145] L Bromham, S Duchêne, X Hua, A M Ritchie, D A Duchêne, and S Y W Ho. Bayesian molecular dating: opening up the black box. *Biological Reviews*, 93(2):1165–1191, 2018.

[146] Z Yang and B Rannala. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution*, 23(1):212–226, 2005.

[147] S Y W Ho and M J Phillips. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Systematic Biology*, 58(3):367–380, 2009.

[148] A J Drummond, O G Pybus, A Rambaut, R Forsberg, and A G Rodrigo. Measurably evolving populations. *Trends in Ecology & Evolution*, 18(9):481–488, 2003.

[149] A Rambaut. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, 16(4):395–399, 2000.

[150] R Bouckaert, J Heled, D Kühnert, T Vaughan, C-H Wu, D Xie, M A Suchard, A Rambaut, and A J Drummond. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4):e1003537, 2014.

[151] M A Suchard, P Lemey, G Baele, D L Ayres, A J Drummond, and A Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1):vey016, 2018.

[152] P Aiewsakun and A Katzourakis. Time-dependent rate phenomenon in viruses. *Journal of Virology*, 90(16):7184–7195, 2016.

[153] Rafael Sanjuán. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. *PLoS Pathogens*, 8(5):e1002685, 2012.

[154] W M Switzer, M Salemi, V Shanmugam, F Gao, M Cong, C Kuiken, V Bhullar, B E Beer, D Vallet, A Gautier-Hion, et al. Ancient co-speciation of simian foamy viruses and primates. *Nature*, 434(7031):376, 2005.

[155] T Sironen, A Vaheri, and A Plyusnin. Molecular evolution of Puumala hantavirus. *Journal of Virology*, 75(23):11803–11810, 2001.

[156] L A Shackelton, C R Parrish, U Truyen, and E C Holmes. High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proceedings of the National Academy of Sciences*, 102(2):379–384, 2005.

[157] L A Shackelton and E C Holmes. Phylogenetic evidence for the rapid evolution of human B19 erythrovirus. *Journal of Virology*, 80(7):3666–3669, 2006.

[158] P Norja, A M Eis-Hübinger, M Söderlund-Venermo, K Hedman, and P Simmonds. Rapid sequence change and geographical spread of human parvovirus B19: comparison of B19 virus evolution in acute and persistent infections. *Journal of Virology*, 82(13):6427–6433, 2008.

[159] T Umemura, Y Tanaka, K Kiyosawa, H J Alter, and J W-K Shih. Observation of positive selection within hypervariable regions of a newly identified DNA virus (SEN virus). *FEBS letters*, 510(3):171–174, 2002.

[160] S Duffy and E C Holmes. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *Journal of Virology*, 82(2):957–965, 2008.

[161] S Y W Ho, R Lanfear, L Bromham, M J Phillips, J Soubrier, A G Rodrigo, and A Cooper. Time-dependent rates of molecular evolution. *Molecular Ecology*, 20(15):3087–3101, 2011.

[162] S Duchêne, E C Holmes, and S Y W Ho. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proceedings of the Royal Society B*, 281(1786):20140732, 2014.

[163] J N Hatwell and P M Sharp. Evolution of human polyomavirus JC. *Journal of General Virology*, 81(5):1191–1200, 2000.

[164] L A Shackelton, A Rambaut, O G Pybus, and E C Holmes. JC virus evolution and its association with human populations. *Journal of Virology*, 80(20):9928–9933, 2006.

[165] A Katzourakis and R J Gifford. Endogenous viral elements in animal genomes. *PLoS Genetics*, 6(11):e1001191, 2010.

[166] M Worobey, G-Z Han, and A Rambaut. Genesis and pathogenesis of the 1918 pandemic H1N1 influenza A virus. *Proceedings of the National Academy of Sciences*, 111(22):8107–8112, 2014.

[167] Z-M Sheng, D S Chertow, X Ambroggio, S McCall, R M Przygodzki, R E Cunningham, O A Maximova, J C Kash, D M Morens, and J K Taubenberger. Autopsy series of 68 cases dying before and during the 1918 influenza pandemic peak. *Proceedings of the National Academy of Sciences*, 108(39):16416–16421, 2011.

[168] T M Tumpey, C F Basler, P V Aguilar, H Zeng, A Solórzano, D E Swayne, N J Cox, J M Katz, J K Taubenberger, P Palese, et al. Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science*, 310(5745):77–80, 2005.

[169] J S Oxford and D Gill. Unanswered questions about the 1918 influenza pandemic: origin, pathology, and the virus itself. *The Lancet Infectious Diseases*, 2018.

[170] J K Taubenberger and J C Kash. Insights on influenza pathogenesis from the grave. *Virus Research*, 162(1-2):2–7, 2011.

[171] D Kobasa, S M Jones, K Shinya, J C Kash, J Copps, H Ebihara, Y Hatta, J H Kim, P Halfmann, M Hatta, et al. Aberrant innate immune response in lethal infection of macaques with the 1918 influenza virus. *Nature*, 445(7125):319, 2007.

[172] A Aswad and A Katzourakis. Paleovirology and virally derived immunity. *Trends in Ecology & Evolution*, 27(11):627–636, 2012.

[173] M Tristem. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *Journal of Virology*, 74(8):3715–3730, 2000.

[174] C Gilbert and C Feschotte. Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biology*, 8(9):e1000495, 2010.

[175] A Margaryan, D Lawson, M Sikora, F Racimo, S Rasmussen, I Moltke, L Cassidy, E Jørsboe, A Ingason, M Pedersen, et al. Population genomics of the Viking world. *bioRxiv*, page 703405, 2019.

[176] H McColl, F Racimo, L Vinner, F Demeter, T Gakuhari, J V Moreno-Mayar, G van Driem, U G Wilken, A Seguin-Orlando, C de la Fuente Castro, et al. The prehistoric peopling of Southeast Asia. *Science*, 361(6397):88–92, 2018.

[177] M Sikora, V Pitulko, V Sousa, M E Allentoft, L Vinner, S Rasmussen, A Margaryan, P de Barros Damgaard, C de la Fuente Castro, G Renaud, et al. The population history of northeastern Siberia since the Pleistocene. *Nature*, 570:182–188, 2019.

[178] E R Jones, G Gonzalez-Fortes, S Connell, V Siska, A Eriksson, R Martiniano, R L McLaughlin, M G Llorente, L M Cassidy, C Gamba, et al. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications*, 6:8912, 2015.

[179] E R Jones, G Zarina, V Moiseyev, E Lightfoot, P R Nigst, A Manica, R Pinhasi, and D G Bradley. The Neolithic transition in the Baltic was not driven by admixture with early European farmers. *Current Biology*, 27(4):576–582, 2017.

[180] G González-Fortes, E R Jones, E Lightfoot, C Bonsall, C Lazar, A Grandal-d'Anglade, M D Garralda, L Drak, V Siska, A Simalcsik, et al. Paleogenomic evidence for multi-generational mixing between Neolithic farmers and Mesolithic hunter-gatherers in the lower Danube basin. *Current Biology*, 27(12):1801–1810, 2017.

[181] M Gallego-Llorente, S Connell, E R Jones, D C Merrett, Y Jeon, A Eriksson, V Siska, C Gamba, C Meiklejohn, R Beyer, et al. The genetics of an early Neolithic pastoralist from the Zagros, Iran. *Scientific Reports*, 6:31326, 2016.

[182] L M Cassidy, R Martiniano, E M Murphy, M D Teasdale, J Mallory, B Hartwell, and D G Bradley. Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proceedings of the National Academy of Sciences*, 113(2):368–373, 2016.

[183] V Siska, E R Jones, S Jeon, Y Bhak, H-M Kim, Y S Cho, H Kim, K Lee, E Veselovskaya, T Balueva, et al. Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. *Science Advances*, 3(2):e1601877, 2017.

[184] C Gamba, E R Jones, M D Teasdale, R L McLaughlin, G Gonzalez-Fortes, V Mattiangeli, L Domboróczki, I Kővári, I Pap, A Anders, et al. Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, 5:5257, 2014.

[185] M G Llorente, E R Jones, A Eriksson, V Siska, K W Arthur, J W Arthur, M C Curtis, J T Stock, M Coltorti, P Pieruccini, et al. Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science*, 350(6262):820–822, 2015.

[186] C L Lai, V Ratziu, M-F Yuen, and T Poynard. Viral hepatitis B. *The Lancet*, 362(9401):2089–2094, 2003.

[187] A Schweitzer, J Horn, R T Mikolajczyk, G Krause, and J J Ott. Estimations of worldwide prevalence of chronic hepatitis B virus infection: a systematic review of data published between 1965 and 2013. *The Lancet*, 386(10003):1546–1555, 2015.

[188] M V Murhekar, K M Murhekar, and S C Sehgal. Epidemiology of hepatitis B virus infection among the tribes of andaman and nicobar islands, india. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 102(8):729–734, 2008.

[189] M V Murhekar, K M Murhekar, V A Arankalle, and S C Sehgal. Epidemiology of hepatitis b infection among the Nicobarese–a mongoloid tribe of the Andaman and Nicobar Islands, India. *Epidemiology & Infection*, 128(3):465–471, 2002.

[190] WHO. Hepatitis B Fact Sheet. `http://www.who.int/mediacentre/factsheets/fs204/en/`, 2017. Accessed: 2017-12-15.

[191] S Locarnini, M Littlejohn, M N Aziz, and L Yuen. Possible origins and evolution of the hepatitis B virus (HBV). *Seminars in Cancer Biology*, 23(6):561–575, 2013.

[192] M Littlejohn, S Locarnini, and L Yuen. Origins and evolution of hepatitis B virus and hepatitis D virus. *Cold Spring Harbor Perspectives in Medicine*, 6(1):a021360, 2016.

[193] A Kramvis. Genotypes and genetic variability of hepatitis B virus. *Intervirology*, 57(3-4):141–150, 2014.

[194] C Hannoun, P Horal, and M Lindh. Long-term mutation rates in the hepatitis B virus genome. *Journal of General Virology*, 81(1):75–83, 2000.

[195] G Zehender, E Ebranati, E Gabanelli, C Sorrentino, A L Presti, E Tanzi, M Ciccozzi, and M Galli. Enigmatic origin of hepatitis B virus: an ancient travelling companion or a recent encounter? *World Journal of Gastroenterology*, 20(24):7622, 2014.

[196] A Kramvis, K Arakawa, M C Yu, R Nogueira, D O Stram, and M C Kew. Relationship of serological subtype, basic core promoter and precore mutations to genotypes/subgenotypes of hepatitis B virus. *Journal of Medical Virology*, 80(1):27–46, 2008.

[197] D M MacDonald, E C Holmes, J C M Lewis, and P Simmonds. Detection of hepatitis B virus infection in wild-born chimpanzees (Pan troglodytes verus): phylogenetic relationships with human and other primate genotypes. *Journal of Virology*, 74(9):4253–4257, 2000.

[198] A H Reid, T G Fanning, J V Hultin, and J K Taubenberger. Origin and evolution of the 1918 'Spanish' influenza virus hemagglutinin gene. *Proceedings of the National Academy of Sciences*, 96(4):1651–1656, 1999.

[199] W W Bond, M S Favero, N J Petersen, C R Gravelle, J W Ebert, and J E Maynard. Survival of hepatitis B virus after drying and storage for one week. *The Lancet*, 317(8219):550–551, 1981.

[200] J F Drexler, A Geipel, A König, V M Corman, D van Riel, L M Leijten, C M Bremer, A Rasche, V M Cottontail, G D Maganga, et al. Bats carry pathogenic hepadnaviruses antigenically related to hepatitis B virus and capable of infecting human hepatocytes. *Proceedings of the National Academy of Sciences*, 110(40):16151–16156, 2013.

[201] L Y Geer, A Marchler-Bauer, R C Geer, L Han, J He, S He, C Liu, W Shi, and S H Bryant. The NCBI biosystems database. *Nucleic Acids Research*, 38(suppl_1):D492–D496, 2009.

[202] T G Bell, M Yousif, and A Kramvis. Bioinformatic curation and alignment of genotyped hepatitis B virus (HBV) sequence data from the genbank public database. *SpringerPlus*, 5(1):1896, 2016.

[203] C Bronk Ramsey. Bayesian analysis of radiocarbon dates. *Radiocarbon*, 51:337–360, 2009.

[204] P J Reimer, E Bard, A Bayliss, J W Beck, P G Blackwell, C B Ramsey, C E Buck, H Cheng, R L Edwards, M Friedrich, et al. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon*, 55(4):1869–1887, 2013.

[205] S Lindgreen. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Research Notes*, 5(1):337, 2012.

[206] H Li and R Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[207] B Buchfink, C Xie, and D H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, 2015.

[208] C Drosten, M Weber, E Seifried, and W K Roth. Evaluation of a new PCR assay with competitive internal control sequence for blood donor screening. *Transfusion*, 40(6):718–724, 2000.

[209] H Jónsson, A Ginolhac, M Schubert, P L F Johnson, and L Orlando. mapDamage2.0: fast approximate bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, page btt193, 2013.

[210] A W Briggs, U Stenzel, M Meyer, J Krause, M Kircher, and S Pääbo. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research*, 38(6):e87–e87, 2009.

[211] M Kearse, R Moir, A Wilson, S Stones-Havas, M Cheung, S Sturrock, S Buxton, A Cooper, S Markowitz, C Duran, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012.

[212] S B Needleman and D D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[213] P Simmonds. SSE: a nucleotide and amino acid sequence analysis platform. *BMC research notes*, 5(1):50, 2012.

[214] D P Martin, B Murrell, M Golden, A Khoosal, and B Muhire. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, 1(1), 2015.

[215] M Padidam, S Sawyer, and C M Fauquet. Possible emergence of new geminiviruses by frequent recombination. *Virology*, 265(2):218–225, 1999.

[216] D P Martin, D Posada, K A Crandall, and C Williamson. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Research & Human Retroviruses*, 21(1):98–102, 2005.

[217] J M Smith. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, 34(2):126–129, 1992.

[218] D Posada and K A Crandall. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences*, 98(24):13757–13762, 2001.

[219] M J Gibbs, J S Armstrong, and A J Gibbs. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics*, 16(7):573–582, 2000.

[220] M F Boni, D Posada, and M W Feldman. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics*, 176(2):1035–1047, 2007.

[221] K Katoh and D M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

[222] S Guindon, J-F Dufayard, V Lefort, M Anisimova, W Hordijk, and O Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321, 2010.

[223] D H Huson and D Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, 2006.

[224] R Ronquist and J P Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 2003.

[225] A Rambaut, T T Lam, L M Carvalho, and O G Pybus. Exploring the temporal structure of heterochronous sequences using tempest (formerly Path-O-Gen). *Virus Evolution*, 2(1):vew007, 2016.

[226] E Jones, T Oliphant, and P Peterson. SciPy: open source scientific tools for Python, 2014.

[227] R R Bouckaert and A J Drummond. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evolutionary Biology*, 17(1):42, 2017.

[228] S Duchêne, D Duchêne, E C Holmes, and S Y W Ho. The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Molecular Biology and Evolution*, 32(7):1895–1906, 2015.

[229] R E Kass and A E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[230] A Rambaut, M A Suchard, D Xie, and A J Drummond. Tracer v1.6., 2015.

[231] G Sanchez, V T Holliday, E P Gaines, J Arroyo-Cabrales, N Martínez-Tagüeña, A Kowler, T Lange, G W L Hodgins, S M Mentzer, and I Sanchez-Morales. Human (Clovis)–gomphothere (Cuvieronius sp.) association 13,390 calibrated yBP in sonora, mexico. *Proceedings of the National Academy of Sciences*, 111(30):10972–10977, 2014.

[232] L Bourgeon, A Burke, and T Higham. Earliest Human Presence in North America Dated to the Last Glacial Maximum: New Radiocarbon Dates from Bluefish Caves, Canada. *PLoS One*, 12(1):e0169486, 2017.

[233] I E Andernach, C Nolte, J W Pape, and C P Muller. Slave Trade and Hepatitis B Virus Genotypes and Subgenotypes in Haiti and Africa. *Emerging Infectious Diseases*, 15(1):1222–1228, 2009.

[234] M Kayser, S Brauer, R Cordaux, A Casto, O Lao, L A Zhivotovsky, C Moyse-Faurie, R B Rutledge, W Schiefenhoevel, D Gil, et al. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the pacific. *Molecular Biology and Evolution*, 23(11):2234–2244, 2006.

[235] P Simmonds and S Midgley. Recombination in the genesis and evolution of hepatitis B virus genotypes. *Journal of Virology*, 79(24):15467–15476, 2005.

[236] P Simmonds. Reconstructing the origins of human hepatitis viruses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 356(1411):1013–1026, 2001.

[237] R S Tedder, S L Bissett, R Myers, and S Ijaz. The 'Red Queen' dilemma–running to stay in the same place: reflections on the evolutionary vector of hbv in humans. *Antivir Ther*, 18(3), 2013.

[238] G Zehender, V Svicher, W Gabanelli, E Ebranati, C Veo, A L Presti, E Cella, M Giovanetti, L Bussini, R Salpini, et al. Reliable timescale inference of HBV genotype A origin and phylodynamics. *Infection, Genetics and Evolution*, 32:361–369, 2015.

[239] C Hannoun, A Söderström, G Norkrans, and M Lindh. Phylogeny of African complete genomes reveals a West African genotype A subtype of hepatitis B virus and relatedness between Somali and Asian A1 sequences. *Journal of General Virology*, 86(8):2163–2167, 2005.

[240] J K Pickrell, N Patterson, P-R Loh, M Lipson, B Berger, M Stoneking, B Pakendorf, and D Reich. Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences*, 111(7):2632–2637, 2014.

[241] S Ghosh, P Banerjee, A RoyChoudhury, S Sarkar, A Ghosh, A Santra, S Banerjee, K Das, B Dwibedi, S K Kar, et al. Unique hepatitis B virus subgenotype in a primitive tribal community in eastern india. *Journal of Clinical Microbiology*, 48(11):4063–4071, 2010.

[242] A Kay and F Zoulim. Hepatitis B virus genetic variability and evolution. *Virus Research*, 127(2):164–176, 2007.

[243] G Gallinella. Parvovirus B19 achievements and challenges. *ISRN Virology*, 2013, 2013.

[244] N S Young and K E Brown. Parvovirus B19. *New England Journal of Medicine*, 350(6):586–597, 2004.

[245] K Ozawa, G Kurtzman, and N Young. Replication of the B19 parvovirus in human bone marrow cell cultures. *Science*, 233(4766):883–886, 1986.

[246] M Söderlund, R von Essen, J Haapasaari, U Kiistala, O Kiviluoto, and K Hedman. Persistence of parvovirus B19 DNA in synovial membranes of young patients with and without chronic arthropathy. *The Lancet*, 349(9058):1063–1065, 1997.

[247] L Pyöriä, M Toppinen, E Mäntylä, L Hedman, L-M Aaltonen, M Vihinen-Ranta, T Ilmarinen, M Söderlund-Venermo, K Hedman, and M F Perdomo. Extinct type of human parvovirus B19 persists in tonsillar B cells. *Nature Communications*, 8:14930, 2017.

[248] T Schenk, M Enders, S Pollak, R Hahn, and D Huzly. High prevalence of human parvovirus B19 DNA in myocardial autopsy samples from subjects without myocarditis or dilative cardiomyopathy. *Journal of Clinical Microbiology*, 47(1):106–110, 2009.

[249] S Tanawattanacharoen, R J Falk, J C Jennette, and J B Kopp. Parvovirus B19 DNA in kidney tissue of patients with focal segmental glomerulosclerosis. *American Journal of Kidney Diseases*, 35(6):1166–1174, 2000.

[250] A Gray, L Guillou, J Zufferey, F Rey, A-M Kurt, P Jichlinski, H-J Leisinger, and J Benhattar. Persistence of parvovirus B19 DNA in testis of patients with testicular germ cell tumours. *Journal of General Virology*, 79(3):573–579, 1998.

[251] L A Adamson, L J Fowler, A S Ewald, M J Clare-Salzler, and J A Hobbs. Infection and persistence of erythrovirus B19 in benign and cancerous thyroid tissues. *Journal of Medical Virology*, 86(9):1614–1620, 2014.

[252] P Norja, K Hokynar, L-M Aaltonen, R Chen, A Ranki, E K Partio, O Kiviluoto, I Davidkin, T Leivo, A M Eis-Hübinger, et al. Bioportfolio: lifelong persistence of variant and prototypic erythrovirus DNA genomes in human tissue. *Proceedings of the National Academy of Sciences*, 103(19):7450–7453, 2006.

[253] J Blümel, R Burger, C Drosten, A Gröner, L Gürtler, M Heiden, M Hildebrandt, B Jansen, T Montag-Lessing, R Offergeld, et al. Parvovirus B19–revised. *Transfusion Medicine and Hemotherapy*, 37(6):339, 2010.

[254] K E Brown and P Simmonds. Parvoviruses and blood transfusion. *Transfusion*, 47(10):1745–1750, 2007.

[255] A Servant, S Laperche, F Lallemand, V Marinho, G De Saint Maur, J F Meritet, and A Garbarg-Chenon. Genetic diversity within human

erythroviruses: identification of three genotypes. *Journal of Virology*, 76(18):9124–9134, 2002.

[256] A Parsyan, C Szmaragd, J-P Allain, and D Candotti. Identification and genetic diversity of two human parvovirus B19 genotype 3 subtypes. *Journal of General Virology*, 88(2):428–431, 2007.

[257] N L Toan, A Duechting, P G Kremsner, L H Song, M Ebinger, S Aberle, V Q Binh, D N Duy, J Torresi, R Kandolf, et al. Phylogenetic analysis of human parvovirus B19, indicating two subgroups of genotype 1 in Vietnamese patients. *Journal of General Virology*, 87(10):2941–2949, 2006.

[258] A Ekman, K Hokynar, L Kakkola, K Kantola, L Hedman, H Bondén, M Gessner, C Aberham, P Norja, S Miettinen, et al. Biological and immunological relations among human parvovirus B19 genotypes 1 to 3. *Journal of Virology*, 81(13):6927–6935, 2007.

[259] J M Hübschen, S Mihneva, A F Mentis, F Schneider, Y Aboudy, Z Grossman, H Rudich, K Kasymbekova, I Sarv, J Nedeljkovic, et al. Phylogenetic analysis of human parvovirus B19 sequences from eleven different countries confirms the predominance of genotype 1 and suggests the spread of genotype 3b. *Journal of Clinical Microbiology*, 47(11):3735–3738, 2009.

[260] E D Heegaard and K E Brown. Human parvovirus B19. *Clinical Microbiology Reviews*, 15(3):485–505, 2002.

[261] X Xia. DAMBE6: new tools for microbial genomics, phylogenetics, and molecular evolution. *Journal of Heredity*, 108(4):431–437, 2017.

[262] X Xia and P Lemey. Assessing substitution saturation with DAMBE. *The phylogenetic handbook: a practical approach to DNA and protein phylogeny*, 2:615–630, 2009.

[263] D Martin and E Rybicki. RDP: detection of recombination amongst aligned sequences. *Bioinformatics*, 16(6):562–563, 2000.

[264] H Shen, W Zhang, H Wang, and S Shao. Identification of recombination in the NS1 and VPs genes of Parvovirus B19. *Journal of Medical Virology*, 88(8):1457–1461, 2016.

[265] C Ramsden, E C Holmes, and M A Charleston. Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. *Molecular Biology and Evolution*, 26(1):143–153, 2008.

[266] M S Chapman and M G Rossmann. Single-stranded DNA–protein interactions in canine parvovirus. *Structure*, 3(2):151–162, 1995.

[267] G L Smith and G McFadden. Smallpox: anything to declare? *Nature Reviews Immunology*, 2(7):521, 2002.

[268] C B Coyne. Horsepox: Framing a dual use research of concern debate. *PLoS Pathogens*, 14(10):e1007344, 2018.

[269] N Sklenovská and M Van Ranst. Emergence of monkeypox as the most important orthopoxvirus infection in humans. *Frontiers in Public Health*, 6, 2018.

[270] R C Hendrickson, C Wang, E L Hatcher, and E J Lefkowitz. Orthopoxvirus genome evolution: the role of gene loss. *Viruses*, 2(9):1933–1967, 2010.

[271] S Parker, R Crump, H Hartzler, and R M Buller. Evaluation of Taterapox Virus in Small Animals. *Viruses*, 9(8):203, 2017.

[272] D Baxby. *Jenner's Smallpox Vaccine: The Riddle of Vaccinia Virus and Its Origin*. Heinemann Educational Publishers, 1981.

[273] K A Bratke, A McLysaght, and S Rothenburg. A survey of host range genes in poxvirus genomes. *Infection, Genetics and Evolution*, 14:406–425, 2013.

[274] B Aguado, I P Selmes, and G L Smith. Nucleotide sequence of 21.8 kbp of variola major virus strain Harvey and comparison with vaccinia virus. *Journal of General Virology.*, 73 (Pt 11):2887–2902, 1992.

[275] J B Moore and G L Smith. Steroid hormone synthesis by a vaccinia enzyme: a new type of virus virulence factor. *The EMBO Journal*, 11(9):3490–3490, 1992.

[276] A Alcami and G L Smith. A mechanism for the inhibition of fever by a virus. *Proceedings of the National Academy of Sciences*, 93(20):11029–11034, 1996.

[277] A F Porter, A T Duggan, H N Poinar, and E C Holmes. Comment: Characterization of Two Historic Smallpox Specimens from a Czech Museum. *Viruses*, 9(10), 2017.

[278] C Smithson, J Imbery, and C Upton. Re-Assembly and Analysis of an Ancient Variola Virus Genome. *Viruses*, 9(9), 2017.

[279] Donald R Hopkins. *Princes and Peasants: Smallpox in History*. The University of Chicago Press, Chicago, 1983.

[280] C W Dixon. *Smallpox*. J & A Churchill Ltd, 104 Cloucester Place, London, 1962.

[281] D E Wood and S L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014.

[282] B Langmead and S L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[283] A Kozlov, D Darriba, T Flouri, B Morel, and A Stamatakis. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference, 2018.

[284] P Barbera, A M Kozlov, L Czech, B Morel, D Darriba, T Flouri, and A Stamatakis. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Systematic Biology*, 68(2):365–369, 2018.

[285] L Czech, P Barbera, and A Stamatakis. Methods for automatic reference trees and multilevel phylogenetic placement. *Bioinformatics*, 35(7):1151–1158, 2018.

[286] G Yu, D K Smith, H Zhu, Y Guan, and T T-Y Lam. ggtree: an rpackage for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, 2016.

[287] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[288] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[289] Y Li, D S Carroll, S N Gardner, M C Walsh, E A Vitalis, and I K Damon. On the origin of smallpox: correlating variola phylogenics with historical smallpox records. *Proceedings of the National Academy of Sciences*, 104(40):15787–15792, 2007.

[290] J Liu, S Wennier, L Zhang, and G McFadden. M062 is a host range factor essential for myxoma virus pathogenesis and functions as an antagonist of host SAMD9 in human cells. *Journal of Virology*, 85(7):3270–3282, 2011.

[291] G Sivan, P Ormanoglu, E C Buehler, S E Martin, and B Moss. Identification of Restriction Factors by Human Genome-Wide RNA Interference Screening of

Viral Host Range Mutants Exemplified by Discovery of SAMD9 and WDR6 as Inhibitors of the Vaccinia Virus K1L-C7L- Mutant. *MBio*, 6(4):e01122, 2015.

[292] S Gillard, D Spehner, R Drillien, and A Kirn. Localization and sequence of a vaccinia virus gene required for multiplication in human cells. *Proceedings of the National Academy of Sciences*, 83(15):5573–5577, 1986.

[293] M E Perkus, S J Goebel, S W Davis, G P Johnson, K Limbach, E K Norton, and E Paoletti. Vaccinia virus host range genes. *Virology*, 179(1):276–286, 1990.

[294] J O Langland and B L Jacobs. The role of the PKR-inhibitory genes, E3L and K3L, in determining vaccinia virus host range. *Virology*, 299(1):133–141, 2002.

[295] B Liu, D Panda, J D Mendez-Rios, S Ganesan, L S Wyatt, and B Moss. Identification of Poxvirus Genome Uncoating and DNA Replication Factors with Mutually Redundant Roles. *Journal of Virology*, 92(7), 2018.

[296] A Alcami and G L Smith. A soluble receptor for interleukin-1β encoded by vaccinia virus: A novel mechanism of virus modulation of the host response to infection. *Cell*, 71(1):153–167, 1992.

[297] M Pires de Miranda, P C Reading, D C Tscharke, B J Murphy, and G L Smith. The vaccinia virus kelch-like protein C2L affects calcium-independent adhesion to the extracellular matrix and inflammation in a murine intradermal model. *Journal of General Virology.*, 84(9):2459–2471, 2003.

[298] P M Beard. Vaccinia virus kelch protein A55 is a 64 kda intracellular factor that affects virus-induced cytopathic effect and the outcome of infection in a murine intradermal model. *Journal of General Virology*, 87(6):1521–1529, 2006.

[299] G C Froggatt, G L Smith, and P M Beard. Vaccinia virus gene F3L encodes an intracellular protein that affects the innate immune response. *Journal of General Virology.*, 88(Pt 7):1917–1921, 2007.

[300] G Kochneva, I Kolosova, T Maksyutova, E Ryabchikova, and S Shchelkunov. Effects of deletions of kelch-like genes on cowpox virus biological properties. *Arch. Virol.*, 150(9):1857–1870, 2005.

[301] C Creighton. *A History of Epidemics in Britain: From A. D. 664 to the extinction of plague*. 1891.

[302] R Willan. *Miscellaneous Works of the Late Robert Willan, M.D*. T. Cadell, London, 1821.

[303] J Moore. *The History of the Small Pox*. Longman, Hurst, Rees, Orme, and Brown, Paternaster Row, London, 1815.

[304] E Paschen. Die Pocken. In *Jochmann's Lehrbuch der Infektionskrankheiten*. Springer Verlag, 2 edition, 1924.

[305] A G Carmichael and A M Silverstein. Smallpox in Europe before the Seventeenth Century: Virulent Killer or Benign Disease? *Journal of the History of Medicine and Allied Sciences*, 42(2):147–168, 1987.

[306] A J Drummond, M A Suchard, D Xie, and A Rambaut. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973, 2012.

[307] P Simmonds, P Aiewsakun, and A Katzourakis. Prisoners of war – host adaptation and its constraints on virus evolution. *Nature Reviews Microbiology*, page 1, 2018.

[308] P Lemey, M Salemi, and A-M Vandamme. *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press, 2009.

[309] M Worobey, M Gemmel, D E Teuwen, T Haselkorn, K Kunstman, M Bunce, J-J Muyembe, J-M M Kabongo, R M Kalengayi, E Van Marck, et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*, 455(7213):661, 2008.

[310] J V Membrebe, M A Suchard, A Rambaut, G Baele, and P Lemey. Bayesian inference of evolutionary histories under time-dependent substitution rates. *Molecular Biology and Evolution*, 2019.

[311] J O Wertheim and S L Kosakovsky Pond. Purifying selection can obscure the ancient age of viral lineages. *Molecular Biology and Evolution*, 28(12):3355–3365, 2011.

[312] J Soubrier, M Steel, M S Y Lee, C Der Sarkissian, S Guindon, S Y W Ho, and A Cooper. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Molecular Biology and Evolution*, 29(11):3345–3358, 2012.

[313] S L Fordyce, M-L Kampmann, N L Van Doorn, and M T P Gilbert. Long – term RNA persistence in postmortem contexts. *Investigative genetics*, 4(1):7, 2013.

[314] S L Fordyce, M C Avila-Arcos, M Rasmussen, E Cappellini, J A Romero-Navarro, N Wales, D E Alquezar-Planas, S Penfield, T A Brown, J-P Vielle-Calzada, et al. Deep sequencing of RNA from ancient maize kernels. *PLoS One*, 8(1):e50961, 2013.

[315] O Smith, G Dunshea, M-H S Sinding, S Fedorov, M Germonpre, H Bocherens, and M T P Gilbert. Ancient RNA from Late Pleistocene permafrost and historical canids shows tissue-specific transcriptome survival. *PLoS biology*, 17(7):e3000166, 2019.

[316] O Smith and M T P Gilbert. Ancient RNA. In *Paleogenomics*, pages 53–74. Springer, 2018.

[317] T-N-N Tran, G Aboudharam, D Raoult, and M Drancourt. Beyond ancient microbial DNA: nonnucleotidic biomolecules for paleomicrobiology. *Biotechniques*, 50(6):370–380, 2011.

[318] J Bardill, A C Bader, G Nanibaa'A, D A Bolnick, J A Raff, A Walker, R S Malhi, et al. Advancing the ethics of paleogenomics. *Science*, 360(6387):384–385, 2018.

[319] A M McCollum, Y Li, K Wilkins, K L Karem, W B Davidson, C D Paddock, M G Reynolds, and I K Damon. Poxvirus viability and signatures in historical relics. *Emerging Infectious Diseases*, 20(2):177, 2014.

[320] G Kolata. *Flu: The story of the great influenza pandemic of 1918 and the search for the virus that caused it*. Simon and Schuster, 2001.

[321] S Miller and M J Selgelid. Ethical and philosophical consideration of the dual-use dilemma in the biological sciences. *Science and Engineering Ethics*, 13(4):523–580, 2007.

[322] M H Hjelmsø, S Mollerup, R H Jensen, C Pietroni, O Lukjancenko, A C Schultz, F M Aarestrup, and A J Hansen. Metagenomic analysis of viruses in toilet waste from long distance flights – A new procedure for global infectious disease surveillance. *PLoS One*, 14(1):e0210368, 2019.

[323] M Breitbart, P Salamon, B Andresen, J M Mahaffy, A M Segall, D Mead, F Azam, and F Rohwer. Genomic analysis of uncultured marine viral com-

munities. *Proceedings of the National Academy of Sciences*, 99(22):14250–14255, 2002.

[324] F Lassalle, M Spagnoletti, M Fumagalli, L Shaw, M Dyble, C Walker, M G Thomas, Andrea Bamberg M, and F Balloux. Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Molecular Ecology*, 27(1):182–195, 2018.

[325] Michael R Wilson, Samia N Naccache, Erik Samayoa, Mark Biagtan, Hiba Bashir, Guixia Yu, Shahriar M Salamat, Sneha Somasekar, Scot Federman, Steve Miller, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *New England Journal of Medicine*, 370(25):2408–2417, 2014.

[326] A Zielezinski, S Vinga, J Almeida, and W M Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18(1):186, 2017.

[327] S Flygare, K Simmon, C Miller, Y Qiao, B Kennedy, T Di Sera, E H Graf, K D Tardif, A Kapusta, S Rynearson, et al. Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biology*, 17(1):111, 2016.

[328] S Vinga and J Almeida. Alignment-free sequence comparison – a review. *Bioinformatics*, 19(4):513–523, 2003.

[329] G Rosen, E Garbarine, D Caseiro, R Polikar, and B Sokhansanj. Metagenome Fragment Classification Using N-Mer Frequency Profiles. *Advances in Bioinformatics*, 2008.

[330] H Teeling, J Waldmann, T Lombardot, M Bauer, and F O Glöckner. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, 5(1):163, 2004.

[331] S Nooij, D Schmitz, H Vennema, A Kroneman, and M P G Koopmans. Overview of virus metagenomic classification methods and their biological applications. *Frontiers in Microbiology*, 9:749, 2018.

[332] S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[333] HMMER3. `http://hmmer.org/`. Accessed: 2019-02-22.

[334] M Steinegger, M Meier, M Mirdita, H Voehringer, S J Haunsberger, and J Soeding. HH-suite3 for fast remote homology detection and deep protein annotation. *BioRxiv*, page 560029, 2019.

[335] P Skewes-Cox, T J Sharpton, K S Pollard, and J L DeRisi. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS One*, 9(8):e105067, 2014.

[336] J Huerta-Cepas, D Szklarczyk, K Forslund, H Cook, D Heller, M C Walter, T Rattei, D R Mende, S Sunagawa, M Kuhn, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, 44(D1):D286–D293, 2015.

[337] S Bzhalava, E Hultin, and J Dillner. Extension of the viral ecology in humans using viral profile hidden Markov models. *PLoS One*, 13(1):e0190938, 2018.

[338] D B Kuchibhatla, W A Sherman, B Y W Chung, S Cook, G Schneider, B Eisenhaber, and D G Karlin. Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently 'orphan' viral proteins. *Journal of Virology*, 88(1):10–20, 2014.

[339] A Reyes, J M P Alves, A Durham, and A Gruber. Use of profile hidden Markov models in viral discovery: Current insights. *Advances in Genomics and Genetics*, 7:29–45, 2017.

[340] C Galiez, C N Magnan, F Coste, and P Baldi. VIRALpro: a tool to identify viral capsid and tail sequences. *Bioinformatics*, 32(9):1405–1407, 2016.

[341] S Roux, F Enault, B L Hurwitz, and M B Sullivan. VirSorter: mining viral signal from microbial genomic data. *PeerJ*, 3:e985, 2015.

[342] C Chothia and A M Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4):823–826, 1986.

[343] J Černỳ, B Černá Bolfíková, J Valdés, J, L Grubhoffer, and D Růžek. Evolution of tertiary structure of viral RNA dependent polymerases. *PLoS One*, 9(5):e96070, 2014.

[344] J Černỳ, B C Bolfíková, M de A Paolo, L Grubhoffer, and D Růžek. A deep phylogeny of viral and cellular right-hand polymerases. *Infection, Genetics and Evolution*, 36:275–286, 2015.

[345] J L Herman, C J Challis, A Novák, J Hein, and S C Schmidler. Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. *Molecular Biology and Evolution*, 31(9):2251–2266, 2014.

[346] A J F Griffiths, S R Wessler, R C Lewontin, and S B Carroll. *Introduction to Genetic Analysis, 9th Edition*. W. H. Freeman, 2008.

[347] W Pirovano and J Heringa. *Protein Secondary Structure Prediction*, pages 327–348. Humana Press, Totowa, NJ, 2010.

[348] K Fujiwara, H Toda, and M Ikeguchi. Dependence of alpha-helical and beta-sheet amino acid propensities on the protein fold type. *BMC Structural Biology*, 12(18):1–15, 2012.

[349] C J A Sigrist, E de Catro, L Cerutti, B A Cuche, N Hulo, Bridge A, L Bougueleret, and I Xenarios. New and continuing developments at PROSITE. *Nucleic Acids Research*, 41(D1):D344–D347, 2013.

[350] M O Dayhoff and R M Schwartz. *A model of evolutionary change in proteins*. National Biomedical Research Foundation, 1978.

[351] S Henikoff and Henikoff J G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Science*, 89(22):10915–10919, 1992.

[352] H Dinkel, K Van Roy, S Michael, M Kumar, B Uyar, B Altenberg, V Milchevskaya, M Schneider, H Kühn, A Behrent, S L Dahl, V Damarell, S Diebel, S Kalman, S Klein, A C Knudsen, C Mäder, S Merrill, A Straudt, V Thiel, L Welti, N E Davey, F Diella, and T J Gibson. ELM 2016 - data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Research*, 4(44):D294–D300, 2016.

[353] M A Casey, R Veltkamp, M Goto, M Leman, C Rhodes, and M Slaney. Content-Based Music Information Retrieval: Current Directions and Future Challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008.

[354] A Lerch. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics, First Edition*. Wiley, 2012.

[355] A Wang. An Industrial-Strength Audio Search Algorithm. *ISMIR*, pages 7–13, 2003.

[356] C E Yoon, O O'Reilly, K J Bergen, and G C Beroza. Earthquake detection through computationally efficient similarity search. *Science Advances*, 1(11):e1501057, 2015.

[357] A B R McIntyre, L Rizzardi, M Y Angela, N Alexander, G L Rosen, D J Botkin, S E Stahl, K K John, S L Castro-Wallace, K McGrath, et al. Nanopore sequencing in microgravity. *npj Microgravity*, 2:16035, 2016.

[358] C Smith. 23 Amazing Shazam Statistics and Facts in 2019. `https://expandedramblings.com/index.php/shazam-statistics/`, 2019. Accessed: 2019-05-30.

[359] L Holm and C Sander. Mapping the protein universe. *Science*, 273(5275):595–602, 1996.

[360] L Holm and C Sander. Protein Structure Comparison by Alignment of Distance Matrices. *Journal of Molecular Biology*, 233(1):123–138, 1993.

[361] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[362] H M Berman. The protein data bank: a historical perspective. *Acta Crystallographica Section A: Foundations of Crystallography*, 64(1):88–95, 2008.

[363] A C Steven and P G Spear. Viral glycoproteins and an evolutionary conundrum. *Science*, 313(5784):177–178, 2006.

[364] C H de Trad, Q Fang, and I Cosic. Protein sequence comparison based on the wavelet transform approach. *Protein engineering*, 15(3):193–203, 2002.

[365] E V Koonin, T G Senkevich, and V V Dolja. The ancient Virus World and evolution of cells. *Biology Direct*, 1(1):1, 2006.

[366] L E Kafetzopoulou, K Efthymiadis, K Lewandowski, A Crook, D Carter, J Osborne, E Aarons, R Hewson, J A Hiscox, M W Carroll, et al. Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Eurosurveillance*, 23(50), 2018.

[367] M AlQuraishi. AlphaFold @ CASP13: "What just happened?". `https://moalquraishi.wordpress.com/2018/12/09/alphafold-casp13-what-just-happened/`, 2018.

[368] M AlQuraishi. AlphaFold at CASP13. *Bioinformatics*, 2019.

[369] J D Gardner, G L Smith, D C Tscharke, and P C Reading. Vaccinia virus semaphorin A39R is a 50–55 kDa secreted glycoprotein that affects the outcome of infection in a murine intradermal model. *Journal of General Virology*, 82(9):2083–2093, 2001.

[370] P C Reading, A Khanna, and G L Smith. Vaccinia virus CrmE encodes a soluble and cell surface tumor necrosis factor receptor that contributes to virus virulence. *Virology*, 292(2):285–298, 2002.

[371] K Dai, Y Liu, M Liu, J Xu, W Huang, X Huang, L Liu, Y Wan, Y Hao, and Y Shao. Pathogenicity and immunogenicity of recombinant Tiantan Vaccinia Virus with deleted C12L and A53R genes. *Vaccine*, 26(39):5062–5071, 2008.

[372] R Liu and B Moss. Vaccinia Virus C9 Ankyrin Repeat/F-Box Protein Is a Newly Identified Antagonist of the Type I Interferon-Induced Antiviral State. *Journal of Virology*, 92(9), 2018.

[373] A H Banham and G L Smith. Characterization of vaccinia virus gene B12R. *Journal of General Virology.*, 74 (Pt 12):2807–2812, 1993.

[374] M Holgado, J Falivene, C Maeto, M Amigo, M Pascutti, M Vecchione, A Bruttomesso, G Calamante, M del Médico-Zajac, and M Gherardi. Deletion of A44L, A46R and C12L vaccinia virus genes from the MVA genome improved the vector immunogenicity by modifying the innate immune response generating enhanced and optimized specific T-cell responses. *Viruses*, 8(5):139, 2016.

[375] P C Reading, J B Moore, and G L Smith. Steroid hormone synthesis by vaccinia virus suppresses the inflammatory response to infection. *The Journal of Experimental Medicine*, 197(10):1269–1278, 2003.

[376] K H Martin, D W Grosenbach, C A Franke, and D E Hruby. Identification and analysis of three myristylated vaccinia virus late proteins. *Journal of Virology*, 71(7):5218–5226, 1997.

[377] V I Chernos, T S Vovk, O N Ivanova, T P Antonova, and V N Loparev. Insertion mutants of the vaccinia virus. the effect of inactivating E7R and D8L genes on the biological properties of the virus. *Mol. Gen. Mikrobiol. Virusol.*, (2):30–34, 1993.

[378] N Price, D C Tscharke, M Hollinshead, and G L Smith. Vaccinia virus gene B7R encodes an 18-kda protein that is resident in the endoplasmic reticulum and affects virus virulence. *Virology*, 267(1):65–79, 2000.

[379] D Wilcock, P Traktman, G L Smith, S A Duncan, and W-H Zhang. The vaccinia virus A40R gene product is a nonstructural, type II membrane glycoprotein that is expressed at the cell surface. *Journal of General Virology.*, 80(8):2137–2148, 1999.

[380] D C Tscharke, P C Reading, and G L Smith. Dermal infection with vaccinia virus reveals roles for virus proteins not seen using other inoculation routes. *Journal of General Virology*, 83(8):1977–1986, 2002.

[381] G L Smith, Y S Chan, and S T Howard. Nucleotide sequence of 42 kbp of vaccinia virus strain WR from near the right inverted terminal repeat. *Journal of General Virology*, 72 (Pt 6):1349–1376, 1991.

[382] S N Shchelkunov. Orthopoxvirus genes that mediate disease virulence and host tropism. *Advances in Virology*, 2012:524743, 2012.

[383] J L Shisler and J Xiao-Lu. The Vaccinia Virus K1L Gene Product Inhibits Host NF-κB Activation by Preventing I B Degradation. *Journal of Virology*, 78(7):3553–3560, 2004.

[384] K E Rehm, R F Connor, G J B Jones, K Yimbu, and R L Roper. Vaccinia virus A35R inhibits MHC class II antigen presentation. *Virology*, 397(1):176–186, 2010.

[385] M T Harte, I R Haga, G Maloney, P Gray, P C Reading, N W Bartlett, G L Smith, A Bowie, and L A J O'Neill. The poxvirus protein A52R targets toll-like receptor signaling complexes to suppress host defense. *The Journal of Experimental Medicine*, 197(3):343–351, 2003.

[386] A Bowie, E Kiss-Toth, J A Symons, G L Smith, S K Dower, and L A O'Neill. A46R and A52R from vaccinia virus are antagonists of host IL-1 and toll-like receptor signaling. *Proceedings of the National Academy of Sciences*, 97(18):10162–10167, 2000.

[387] S C Graham, M W Bahar, S Cooray, R A Chen, D M Whalen, N G A Abrescia, D Alderton, R J Owens, D I Stuart, G L Smith, and J M Grimes. Vaccinia Virus Proteins A52 and B14 Share a Bcl-2–Like Fold but Have Evolved to Inhibit NF-κB rather than Apoptosis. *PLoS Pathog.*, 4(8):e1000128, 2008.

[388] A Murcia-Nicolas, G Bolbach, J C Blais, and G Beaud. Identification by mass spectroscopy of three major early proteins associated with virosomes in vaccinia virus-infected cells. *Virus Res.*, 59(1):1–12, 1999.

[389] F A Legrand, P H Verardi, L A Jones, K S Chan, Y Peng, and T D Yilma. Induction of potent humoral and cell-mediated immune responses by attenuated vaccinia virus vectors with deleted serpin genes. *Journal of Virology*, 78(6):2770–2779, 2004.

[390] S Kettle, A Khanna, A Alcami, C Jassoy, R Ehret, and G L Smith. Vaccinia virus serpin B13R (SPI-2) inhibits interleukin-1beta-converting enzyme and protects virus-infected cells from TNF- and Fas-mediated apoptosis, but does not prevent IL-1beta-induced fever. *Journal of General Virology.*, 78(3):677–685, 1997.

[391] S Kettle, N W Blake, K M Law, and G L Smith. Vaccinia virus serpins B13R (SPI-2) and B22R (SPI-1) encode M(r) 38.5 and 40K, intracellular polypeptides that do not affect virus virulence in a murine intranasal model. *Virology*, 206(1):136–147, 1995.

[392] H Berman, K Henrick, and H Nakamura. Announcing the worldwide protein data bank. *Nature Structural and Molecular Biology*, 10(12):980, 2003.

[393] W Kabsch and C Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[394] I Korf, M Yandell, and J Bedell. *BLAST: An Essential Guide to the Basic Local Alignment Search Tool*. OReilly, 2003.

[395] L Holm and Rosenström P. Dali server: conservation mapping in 3D. *Nucleic Acid Research*, 38(suppl 2):W545–W549, 2010.

[396] Schrödinger, L L C. The PyMOL Molecular Graphics System, Version 1.8. 2015.

[397] P Y Chou and G D Fasman. β-Turns in proteins. *Journal of Molecular Biology*, 115(2):135–175, 1977.

[398] J Garnier, J Gibrat, and B Robson. GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence. *Methods in Enzymology*, 266(32):540–553, 1996.

[399] K A Dill and J L MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[400] A M Lesk and C H Chothia. The response of protein structures to amino-acid sequence changes. *Philosophical Transactions of the Royal Society of Lon-*

don A: Mathematical, Physical and Engineering Sciences, 317(1540):345–356, 1986.

[401] B Rost. Prediction in 1D: secondary structure, membrane helices, and accessibility. *Methods of Biochemical Analysis*, 44:559–588, 2003.

[402] P Y Chou and G D Fasman. Prediction of Protein Conformation. *Biochemistry*, 13(2):222–245, 1974.

[403] J Garnier, D J Osguthorpe, and B Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology*, 120(1):97–120, 1978.

[404] W Kabsch and C Sander. How good are predictions of protein secondary structure? *FEBS letters*, 155(2):179–182, 1983.

[405] B Rost and C Sander. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232(2):584–599, 1993.

[406] D T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202, 1999.

[407] G Pollastri, D Przybylski, B Rost, and P Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function, and Bioinformatics*, 47(2):228–235, 2002.

[408] A Kloczkowski, K-L Ting, R L Jernigan, and J Garnier. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins: Structure, Function, and Bioinformatics*, 49(2):154–166, 2002.

[409] Y Huang, B Niu, Y Gao, L Fu, and W Li. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, 2010.

[410] A V Aho and M J Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, 1975.

[411] P Bork and E V Koonin. Protein sequence motifs. *Current Opinion in Structural Biology*, 6(3):366–376, 1996.

[412] J D Forman-Kay and T Mittag. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure*, 21(9):1492–1499, 2013.

[413] N Tokuriki, C J Oldfield, V N Uversky, I N Berezovsky, and D S Tawfik. Do viral proteins possess unique biophysical features? *Trends in Biochemical Sciences*, 34(2):53–59, 2009.

[414] N E Davey, K Van Roey, R J Weatheritt, G Toedt, B Uyar, B Altenberg, A Budd, F Diella, H Dinkel, and T J Gibson. Attributes of short linear motifs. *Molecular BioSystems*, 8(1):268–281, 2012.

[415] C J A Sigrist, L Cerutti, N Hulo, A Gattiker, L Falquet, M Pagni, A Bairoch, and P Bucher. PROSITE: A documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics*, 3(3):265–274, 2002.

[416] S Eddy. Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology*, 22(8):1035–1036, 2004.

[417] X Xia and W-H Li. What amino acid properties affect protein evolution? *Journal of Molecular Evolution*, 47(5):557–564, 1998.